
# TRABAJO FIN DE GRADO:

# CLASIFICACIÓN AUTOMÁTICA DE TEXTO PARA EL SEGUIMIENTO DE CAMPAÑAS ELECTORALES EN REDES SOCIALES

AUTORA: CRISTINA GONZÁLEZ RUBIO

TUTOR: JULIO VILLENA ROMÁN.

TITULACIÓN: GRADO EN INGENIERÍA DE SISTEMAS DE COMUNICACIONES.

FECHA: 8 de octubre de 2015

# AGRADECIMIENTOS

En primer lugar, quiero agradecer a mi tutor Julio Villena Román toda la atención, amabilidad y dedicación que he recibido de su parte.

Gracias a mis compañeros de clase por haber compartido esta etapa de mi vida, esas horas interminables de laboratorio, y como olvidar las horas y horas en la biblioteca. Me llevo no solo compañeros de clase, sino verdaderos amigos.

Gracias a mis amigos, por el seguimiento en todo momento de mi carrera y sus ánimos.

Gracias a mis padres y hermano por su incondicional apoyo. Gracias por vuestros ánimos en mis momentos más bajos y por vuestra comprensión. Sin vosotros no lo hubiera logrado.

Gracias a mi familia política que me ayudaron en los malos momentos y celebramos los buenos momentos. Gracias Ceci, Alberto, Luci, Juani y Paco. Gracias por formar parte de esto.

Gracias a ti, Daniel, por tu confianza, por tu apoyo, por estar siempre a mi lado animándome. Este proyecto es para ti. Todo esto es gracias a ti.

# Resumen

En la actualidad existe una inmensa cantidad de información disponible en formato electrónico. Toda esta información es improductiva si no se dispone de mecanismos apropiados para su acceso, clasificación y análisis.

En el presente trabajo, se han desarrollado soluciones específicas de clasificación automática de textos en el ámbito de la política. Concretamente se analizan mensajes, de la popular red social Twitter (una de las redes sociales con mayor aceptación entre los usuarios), de los principales partidos políticos y sus principales representantes, en plena campaña electoral. El hecho de desarrollar un sistema automático de clasificación evitará la intervención humana aumentando así la rapidez con que se pueden procesar las opiniones políticas. Esto conllevará a la reducción sustancial de costes frente a una clasificación manual.

La decisión de implementar un clasificador, empleando como documentos de trabajo las diferentes opiniones políticas, se antoja muy conveniente para analizar un sistema capaz de detectar las diferentes cuestiones tratadas durante la campaña electoral.

El tiempo que dura la campaña es uno de los periodos óptimos para la recopilación de opiniones políticas, ya que es el momento en el que todos los partidos y representantes políticos de las distintas fuerzas intentan llamar la atención del ciudadano para captar su voto.

A lo largo del Trabajo Fin de Grado, rescatamos por medio de la API de Twitter, dos grupos de opiniones políticas. El primero de los grupos es utilizado en la fase de entrenamiento, donde se llevará a cabo la implementación del clasificador y la definición de las clases, en las cuales se clasificarán los tweets. Previo a la definición de clases, será necesario definir unas reglas de clasificación. Para la realización de esta fase se recopilan 300 tweets.

Se ofrece un estudio detallado de las actuales técnicas y líneas de investigación, sobre la clasificación automático de texto.

Los tweets se clasifican en una o varias clases, en función de su contenido. En la Figura 1 puede verse un ejemplo de la clasificación de tres tweets. El primer y segundo tweet son clasificados como *Esperanza*, y el tercer tweet es clasificado como *Esperanza, Desigualdad* y como *Cambio.*

| Text | Classes |
|---|---|
| RT @SobresEnB: Después d ver la explotación d trabajos sociales en un pueblo de Valencia ¿Q esperar? @RamonEspinar @Pablo_Iglesias_ http://… | Esperanza (100%) esperar |
| Después d ver la explotación d trabajos sociales en un pueblo de Valencia ¿Q esperar? @RamonEspinar @Pablo_Iglesias_ http://t.co/0zrLm5s1Q1 | Esperanza (100%) esperar |
| RT @Pablo_Iglesias_: Medidas, políticas, coraje, cambio de actitud, eso se espera de un gobierno que lucha contra la violencia machista htt… | Esperanza (100%) espera<br>Desigualdad (100%) violencia<br>Cambio (100%) cambio |

**Figura 1: Ejemplo de clasificación de tweets.**

El segundo de los grupos de opiniones, es para la fase de validación del clasificador. Una vez obtenidos los tweets a clasificar se realiza la exportación de todos los datos a un archivo Excel, ya que de esta manera y aplicando los filtros correspondientes se recopilan los resultados de una forma más sencilla y visual para su posterior validación. Con la totalidad de los datos en el archivo Excel se obtiene la matriz de confusión, elemento fundamental para el análisis del clasificador.

Para esta fase de validación se cuenta con un total de 1200 tweets para su estudio, entre partidos políticos y sus representantes. Posteriormente se mostrarán los resultados obtenidos de la validación del clasificador, extrayendo las conclusiones más representativas del clasificador implementado, y se revelarán las dificultades encontradas a la hora de implementar dicho clasificador automático de texto, en el ámbito de la política.

Se han obtenido unos resultados aceptables en el caso práctico realizado, con unos valores de precisión y cobertura superiores al 60% para las clases definidas. Permite hacerse una idea de la viabilidad de estos sistemas.

Además se incluye un presupuesto sobre el Trabajo Fin de Grado, donde se detallan los costes tanto directos como indirectos. La planificación del trabajo se detalla mediante un diagrama de Gantt.

Por último, se expondrán una serie de trabajos futuros, que podrían mejorar sensiblemente la precisión y cobertura del clasificador implementado.

# Contenido

# Índice de figuras

# Índice de tablas

# 1. Introducción

## 1.1. Motivación

Hoy en día, podemos acceder a gran cantidad de información. Esto es gracias a los avances en la tecnología de la información y comunicación (TIC), que ofrecen a la sociedad y a las empresas una gran cantidad de oportunidades y retos.

Hasta hace una década, la gran mayoría de los datos producidos en el mundo eran resultado de procesos científicos, industriales y administrativos. Pero la explosión de las tecnologías móviles y la popularización de los servicios sociales de la Web 2.0 han cambiado esto de manera radical: actualmente el principal agente de la explosión de datos es la actividad cotidiana de millones de ciudadanos.

Hoy no somos solamente consumidores de datos. Las plataformas sociales de Internet construyen un perfil extremadamente detallado de nuestras preferencias y nos convierten en un producto. Nuestros datos son la mercancía con la que comercian los gestores de información (data brokers) y un componente esencial del modelo económico que sostiene a Internet. Esta recogida sistemática de datos sobre nuestra vida personal es uno de los factores que hacen posible el estado de vigilancia masiva. Tanto las grandes empresas como las PYMEs, están utilizando las redes sociales como elemento de posicionamiento y venta.

Es importante para las empresas, o en el caso que abarca este Trabajo Fin de Grado para los partidos políticos, saber el tipo de conversaciones que se dan en su perfil, o el tipo de seguidores que tiene en su canal de Twitter. Por ello, cada vez se demandan más las tareas de análisis de gestión y de monitorización.

Ya sea realizando búsquedas en Google, subiendo vídeos a YouTube, actualizando Twitter o aceptando solicitudes en Facebook, nuestras acciones producen una gran cantidad de huellas digitales en las que quedan capturados nuestros deseos, miedos y esperanzas. Por este motivo, actualmente se están utilizando técnicas como el *sentiment analysis,* traducido al castellano como análisis de sentimientos, para intentar determinar nuestras preferencias colectivas a la hora de comprar un producto u opinar sobre una decisión política.

Esta producción de datos en cantidades masivas es uno de los hechos fundamentales de nuestro tiempo. Para definir este fenómeno, se acuñó el término *Big Data*, proveniente del mundo anglosajón, cuya traducción es datos masivos.

Dado que se maneja una gran cantidad de datos, se deben ordenar mediante unas reglas de clasificación para su mejor comprensión, procesado y análisis.

La clasificación es un concepto muy común entre los profesionales que se dedican a la documentación. Consiste en organizar los documentos de algún modo que permita después su mejor recuperación y posterior análisis. Con la creciente disponibilidad de información en formato electrónico a través de blogs, foros, páginas de opinión, redes sociales, etc.,

susceptibles de ser procesados de manera automática, surge la posibilidad de abordar la clasificación de documentos de manera automática.

Mientras que en los últimos quince años el coste de almacenar información digital ha disminuido enormemente, el número de dispositivos que captan, producen, procesan y transmiten datos se ha multiplicado de manera exponencial. Tener acceso a más datos no es solo una cuestión de volumen; a partir de un umbral determinado, es posible hacer las cosas de otra manera. Las inmensas masas de información que producen las organizaciones científicas, empresariales y gubernamentales contienen grandes bolsas de conocimiento valioso que pueden ser capturadas si aprendemos a detectarlas, extraerlas y leerlas, es decir, a procesarlas, visualizarlas y analizarlas.

La revolución de los datos masivos ha traído consigo un conjunto de nuevas metodologías y técnicas de análisis y gestión de la información, así como profesiones emergentes: del data scienstist al analista de datos y el experto en visualización de la información.

Debido a esta gran cantidad de información surge la minería de datos (termino adoptado a partir del anglicismo *data mining*). Es un conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.

Cada vez es más común el desarrollo de actividades asociadas al tratamiento de **grandes volúmenes de datos**, y en particular para realizar análisis estadísticos y computacionales complejos de los mismos, con el objetivo de obtener **resultados aplicables en identificación de patrones, predicción, simulación u optimización**. Estos resultados deben suponer el punto de partida básico para la mejora de la eficiencia en empresas y organizaciones.

Situar la cultura de los datos en el centro de la toma de decisiones y de nuestra manera de interpretar el mundo abre muchas posibilidades, pero también implica numerosos riesgos. El principal peligro del datacentrismo es que fomente la idea de que en los datos se encuentra la respuesta a cualquier problema y que nuestra sociedad puede prescindir de mecanismos más imperfectos y desordenados, basados en la política y la negociación.

Preservar valores como la subjetividad y la ambigüedad es especialmente importante en un momento en que es fácil pensar que todas las soluciones son computables y se encuentran dentro de un servidor, almacenadas en un Data Center.

Como ejemplo ilustrativo de esta demanda, el gobierno francés ha estimado que las necesidades de profesionales con un perfil orientado al tratamiento de grandes volúmenes de datos durante los próximos cinco años, y para el conjunto de la UE, excederán de las 300.000 personas.

En España, se prevé que el empleo de estas técnicas aumente al menos en un 300% en los próximos 3 años. Un 11% de las empresas españolas medianas y grandes han mostrado su interés por aplicar dichas técnicas, que en este momento solo se emplean en un 5% de las mismas.

## 1.2. Objetivos

El objetivo de este Trabajo Fin de Grado es la implementación y análisis de un sistema de clasificación automática de textos, concretamente de opiniones de la clase política enmarcadas en periodo de campaña electoral. La plataforma elegida para la recopilación de opiniones políticas, es la red social Twitter.

Los tweets serán clasificados, en una de las categorías definidas en el sistema, pudiéndose clasificar en más de una categoría simultáneamente.

Los principales objetivos de este proyecto se pueden resumir en los siguientes puntos:

- Recopilación de tweets mediante la API de Twitter, para la implementación y posterior evaluación del clasificador automático de texto.
- Ser capaces de extraer la información relevante contenida en los tweets de opinión política, para obtener una representación estructurada de los mismos que facilite su procesado y análisis.
- Lograr que el conocimiento adquirido a partir de tweets ya categorizados, permita desarrollar un sistema de clasificación automática válido y eficiente ante una consulta del usuario.
- Entender y analizar los problemas que se puedan derivar del empleo de tweets de opinión política, o tweets con una carga afectiva significativa, a la hora de implementar un clasificador automático de documentos.

## 1.3. Organización del documento

Este documento se estructura en varios apartados detallados a continuación mediante una breve descripción:

1. Resumen.

   En este apartado se realiza una breve explicación acerca del contenido del trabajo y las razones sobre las que se apoya la idea de la realización de un clasificador automático de texto. También se enumeran razones por la cuales se ha pensado que es el momento óptimo para basar el clasificador en opiniones políticas

2. Introducción.

   Se detallan las motivaciones y los objetivos planteados para la realización del trabajo. En este apartado también se muestra la organización del documento.

3. Estado del Arte.

   En este apartado se realiza un repaso sobre los conocimientos previos que dan paso a la realización del trabajo. También se detalla la evolución de la tecnología de los clasificadores de texto, así como una reseña sobre la evolución de las redes sociales,

centrando la atención en la red social que se ha elegido para la recopilación, clasificación y análisis de diferentes opiniones políticas: Twitter.

4. Marco regulador.

Se hace repaso de la ley vigente para asegurar la legalidad del documento, sin incurrir en posibles delitos que afecten a la política de privacidad de datos.

5. Captura de datos.

Fase del trabajo en el que se procede a la recopilación de todos los tweets que se van a utilizar tanto en la fase de entrenamiento (300 tweets) como en la fase de validación (1200 tweets) del clasificador. Los tweets recopilados para la fase de entrenamiento es de un orden mucho menor que los recopilados para la fase de validación. En ningún caso se usan los tweets de una fase para la consecución de la otra.

6. Fase de entrenamiento.

Apartado dedicado a la definición de las reglas y el etiquetado de las clases. Tras estas fases se procede a la implementación del clasificador. El número de tweets recopilados para esta fase es de 300.

7. Resultados.

En este apartado se lleva a cabo la validación del clasificador implementado anteriormente, a través de la construcción de la matriz de confusión. Para la obtención de la matriz de confusión se han evaluado un total de 1200 tweets. Para la mejor comprensión de la fase de validación, se explican los diferentes valores que se obtienen a través de la matriz de confusión y su utilidad para la validación del clasificador. Se detallan las gráficas y los datos que se obtienen en la fase de validación y en los que nos basamos para llegar a las conclusiones sobre la viabilidad del clasificador.

8. Conclusiones.

Se evalúan los resultados identificando las clases con mayor precisión y cobertura, y analizando los puntos fuertes del clasificador y los aspectos a mejorar, basándose en los resultados anteriores.

9. Trabajos futuros.

En el último apartado del trabajo se presentan posibles vías de mejora del clasificador y las posibles líneas de investigación acerca de la clasificación automática de textos.

# 2. Estado del arte

## 2.1. Introducción

Diariamente se trabaja con grandes cantidades de documentos escritos que tienen que ser clasificados, seleccionados o distribuidos de maneras diferentes para poder ser tratados adecuadamente. Realizar esta tarea manualmente requiere una enorme cantidad de tiempo y esfuerzo.

La clasificación automática de textos puede definirse como la acción ejecutada por un sistema artificial sobre un conjunto de elementos para ordenarlos en clases o categorías. Los elementos a clasificar pueden ser de cualquier tipo (TXT, Word, PDF, etc.).

La clasificación automática de textos es una de las áreas de investigación que ha cobrado mayor importancia en los últimos años debido, en parte, a los grandes volúmenes de textos digitales que se almacenan en bases de datos empresariales, páginas web, comentarios en foros y redes sociales.

Los principales beneficios que se obtienen con la clasificación automática de textos son los siguientes:

- Reducción sustancial de costes frente a una clasificación manual y mejora de la productividad de los distintos departamentos de una empresa.
- Sistemas de alto rendimiento capaces de procesar grandes volúmenes de texto en tiempo real.
- Diseño a medida de las necesidades del cliente.
- Agilidad en la toma de decisiones y mejora de la planificación.


**Figura 2: Proceso de clasificación automática de texto.**

Por ello, poder organizar la información de forma automática ha pasado a ser una tarea de vital importancia y llevar a cabo una gestión eficiente de la información se ha convertido en algo imprescindible. Por este motivo cada vez son más necesarias las herramientas que puedan automatizar esta clasificación.

## 2.2. Conocimientos previos

### 2.2.1. Evolución de la clasificación automática de textos

Es en la década de los años 60 cuando se presentan los primeros clasificadores automáticos de texto [28]. Desde estas fechas hasta la década de los 80 y principios de los 90, la clasificación de textos se llevaba a cabo mediante un proceso manual que extraía el conocimiento del experto y lo representaba mediante reglas por medio de técnicas de ingeniería del conocimiento. Estas reglas se construyen como:

$$if\ <condición_i>\ then\ <clase_j>$$

donde, si el texto a clasificar satisface la condición i-ésima entonces es clasificado en la clase o categoría j-ésima. Un ejemplo de este tipo de clasificadores es el Sistema Construe [24], construido por ***Carnegie Group*** para la agencia de noticias Reuters. La principal desventaja de este enfoque radica en la dificultad de extraer el conocimiento del experto, lo que provoca, que dichos clasificadores no sean portables, porque las reglas obtenidas son específicas del problema y del dominio; y difícilmente mantenibles, porque pueden surgir nuevas reglas que deben ser definidas por el experto.

Es en la década de los 90, cuando el paradigma de la máquina que aprende [29], emerge como un nuevo enfoque de clasificación que atrae el interés de diferentes investigadores. En dicho enfoque aparece un proceso que se denomina *proceso general inductivo*, que construye de forma automática un clasificador por aprendizaje a partir de un conjunto de textos previamente clasificados. Para ello, este proceso extrae las características que debe tener un texto, desde unos ejemplos de entrenamiento dados por un experto, para pertenecer a una clase. Por lo tanto, con este enfoque el esfuerzo del ingeniero no se dirige hacia la construcción de un clasificador, sino que se dirige hacia la confección de un proceso automático de construcción de clasificadores. De manera que, si el conjunto original de clases se actualiza o el sistema es portado a un dominio diferente, solamente es necesario realizar un nuevo entrenamiento a partir del nuevo conjunto de textos.

Las principales ventajas que presenta este enfoque son:

- *Efectividad*, ya que no es necesario que un experto defina las reglas de clasificación.
- *Independencia del dominio* de los textos a clasificar.

La gran mayoría de los clasificadores de textos por aprendizaje se basan en métodos de inducción probabilísticos, esencialmente cuantitativos (numéricos), lo que conlleva una difícil interpretación de los resultados. Otra clase de clasificadores que han experimentado un gran auge en los últimos años, son los simbólicos. Estos se basan en la localización y posterior clasificación de los patrones más representativos del texto y determinantes de cada categoría. Los clasificadores construidos bajo este nuevo paradigma están alcanzando resultados que hacen de la clasificación automática por aprendizaje una alternativa cualitativa y comercialmente viable respecto a los clasificadores tradicionales.

### 2.2.2. Técnicas de aprendizaje

La clasificación automática de textos ha estado ligada históricamente al desarrollo de aprendizaje automático o computacional (machine learning), una línea de la Inteligencia Artificial y la Inteligencia Computacional que se basa en el desarrollo de un conjunto de algoritmos, que "aprenden" o reconocen patrones recurrentes en cada clase a partir de un gran volumen textos de entrada, previamente clasificados por humanos.

Este conjunto de algoritmos, junto con técnicas y sistemas, son capaces de asignar un documento a una o varias categorías, según su afinidad temática. Las técnicas de aprendizaje que se emplean son:

- Aprendizaje automático (ML: Machine Learning)
- Procesamiento del lenguaje natural (NLP: Natural Language Processing)

El **Aprendizaje Automático [34]** (AA, o **Machine Learning,** por su nombre en inglés) es la rama de la Inteligencia Artificial que tiene como objetivo desarrollar técnicas que permitan a las computadoras aprender. Se trata de crear algoritmos capaces de generalizar comportamientos y reconocer patrones a partir de una información suministrada en forma de ejemplos. Es, por lo tanto, un proceso de **inducción del conocimiento**, es decir, un método que permite obtener por generalización un enunciado general a partir de enunciados que describen casos particulares.

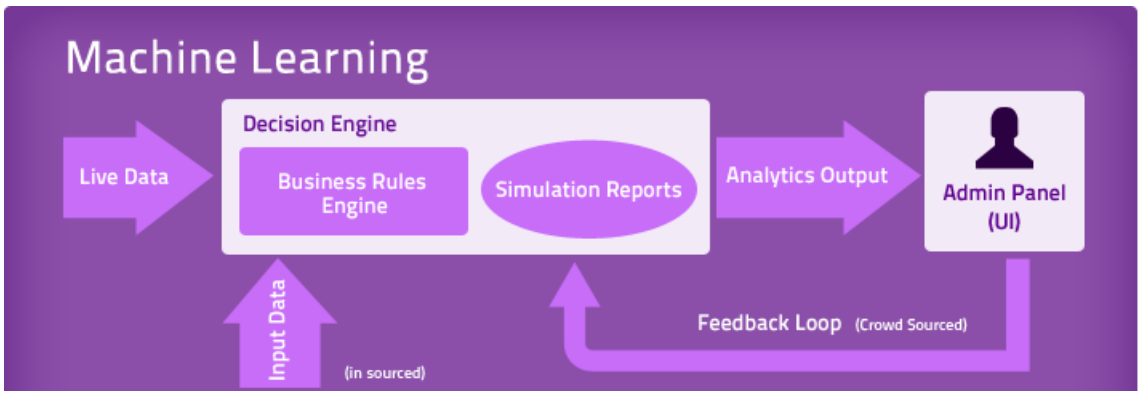



Figura 3. Esquema de aprendizaje automático [14].

Cuando se han observado todos los casos particulares la inducción se considera **completa**, por lo que la generalización a la que da lugar se considera **válida** [34]. No obstante, en la mayoría de los casos se antoja imposible obtener una inducción completa, por lo que el enunciado a que da lugar queda sometido a un cierto grado de incertidumbre, y en consecuencia no se puede considerar como un esquema de inferencia formalmente válido ni se puede justificar empíricamente. En la mayoría de las ocasiones el campo de actuación del aprendizaje automático se solapa con el de **Data Mining**, ya que las dos disciplinas están enfocadas en el análisis de datos, sin embargo el aprendizaje automático se centra más en el

estudio de la complejidad computacional de los problemas con la intención de hacerlos factibles desde el punto de vista práctico, no únicamente teórico. [34]

Dando una explicación más coloquial, se puede decir que puede ser que una de las tareas del AA es intentar extraer conocimiento sobre algunas propiedades no observadas de un objeto basándose en las propiedades que sí han sido observadas de ese mismo objeto (o incluso de propiedades observadas en otros objetos similares)... o, dicho de otra manera, **predecir** un comportamiento futuro a partir de un hecho que ha ocurrido en el pasado. [34] Por poner un ejemplo sencillo: predecir si a un cliente le va a gustar un producto determinado sin que lo haya probado, basándose en las opiniones que ha dado en base a la experiencia con otro producto similar que sí ha probado.

En cualquier caso, como el tema del que estamos hablando está relacionado con el aprendizaje, lo primero que hemos de preguntarnos es: **¿Qué entendemos por aprender?** y, ya que queremos dar metodologías generales para producir un aprendizaje de forma automática, una vez que fijemos este concepto habremos de dar métodos para medir el grado de éxito/fracaso de un aprendizaje. En cualquier caso, ya que estamos trasladando un concepto intuitivo y que usamos normalmente en la vida diaria a un contexto computacional, ha de tenerse en cuenta que todas las definiciones que demos de aprendizaje desde un punto de vista computacional, así como las diversas formas de medirlo, estarán íntimamente relacionadas con contextos muy concretos y posiblemente lejos de lo que intuitivamente, y de forma general, entendemos por aprendizaje. [34]

Una definición relativamente general de **aprendizaje** dentro del contexto humano podría ser la siguiente: *proceso a través del cual se adquieren o modifican habilidades, destrezas, conocimientos, conductas o valores como resultado del estudio, la experiencia, la instrucción, el razonamiento y la observación*. De esta definición es importante hacer notar que el aprendizaje debe producirse a partir de la experiencia con el entorno, no se considera aprendizaje toda aquella habilidad o conocimiento que sean innatos en el individuo o que se adquieran como resultado del crecimiento natural de este. Siguiendo un esquema similar, en el AA vamos a considerar aprendizaje a aquello que la máquina pueda **aprender a partir de la experiencia**, no a partir del reconocimiento de patrones programados a priori. Por tanto, una tarea central de cómo aplicar esta definición al contexto de la computación va a consistir en alimentar la experiencia de la máquina por medio de objetos con los que entrenarse (ejemplos) para, posteriormente, aplicar los patrones que haya reconocido sobre otros objetos distintos (en un sistema de recomendación de productos, un ejemplo sería un par particular cliente/producto, junto con la información acerca de la valoración que aquel haya hecho de este). [34]

El **Procesamiento de Lenguaje Natural** (PLN ó **Natural Language Processing**) [22] estudia los problemas inherentes al procesamiento y manipulación de lenguajes naturales, haciendo uso de ordenadores. Pretende adquirir conocimiento sobre el modo en que los humanos entienden y utilizan el lenguaje, de tal forma que se pueda llevar a cabo el desarrollo de herramientas y técnicas para conseguir que los ordenadores puedan entenderlo y manipularlo. Sus fundamentos residen en un conjunto muy amplio de disciplinas: ciencias de la información

y los computadores, lingüística, matemáticas, ingeniería eléctrica y electrónica, inteligencia artificial y robótica, psicología, etc.

Existe un gran número de aplicaciones donde el PLN resulta de gran utilidad (traducción máquina, procesamiento y resumen de textos escritos en lenguaje natural, interfaces de usuario, reconocimiento de voz, etc.). Para el diseño de la función de clasificación se pueden emplear diferentes técnicas de aprendizaje, debiendo disponer para ello de un conjunto de documentos (conjunto de entrenamiento, véase Figura 4), que previamente han sido clasificados dentro de una determinada categoría. Estos algoritmos de aprendizaje o entrenamiento requieren una representación estructurada de los documentos. La más empleada es la basada en el modelo de espacio vectorial, donde cada documento se transforma en un vector de palabras clave a las que se les asigna un peso en función de la importancia o relevancia que estas representen dentro del documento. Una vez que el clasificador ha sido entrenado con el correspondiente grupo de textos, su efectividad se evalúa comparando las categorías que ha asignado a los documentos del set de prueba con las que estos ya tenían asignadas. Este esquema permite alcanzar una precisión comparable a la obtenida por expertos humanos, reduciendo así los costes de mano de obra [22].

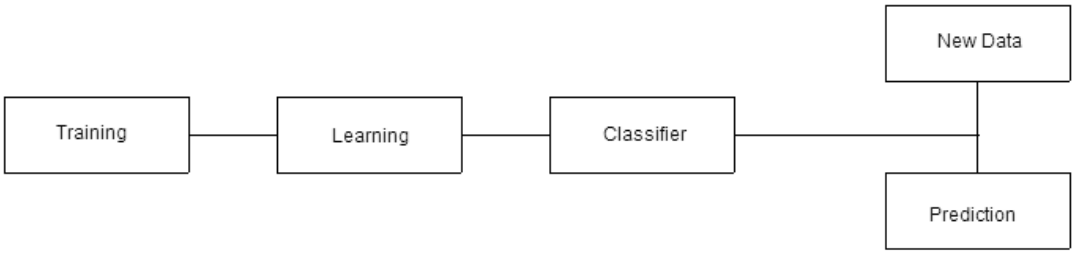



**Figura 4: Clasificación de texto dando un conjunto de entrenamiento.**

Algunos ejemplos de los entornos en los que se emplea la clasificación automática son: indexación automática de textos, filtrado de textos, clasificación de páginas Web, filtrado de correos electrónicos (spam), o clasificación de noticias.

### 2.2.3. Tipos de clasificadores según sus características

Como se puede ver en la siguiente figura, los clasificadores se clasifican principalmente en cuatro categorías:

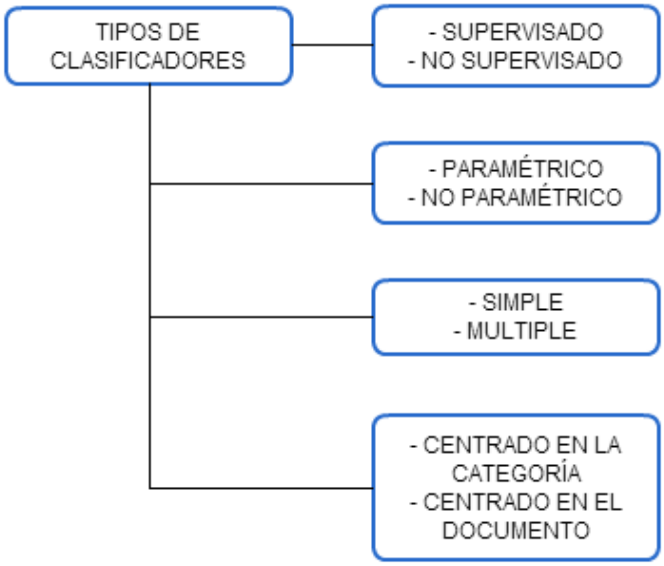



**Figura 5: Tipos de clasificadores según sus características.**

### *Clasificación supervisada*

El aprendizaje supervisado es una técnica para deducir una función a partir de datos de entrenamiento [4]. Los datos de entrenamiento consisten en pares de objetos (normalmente vectores): una componente del par son los datos de entrada y el otro, los resultados deseados. La salida de la función puede ser un valor numérico (como en los problemas de regresión) o una etiqueta de clase (como en los de clasificación). El objetivo del aprendizaje supervisado es el de crear una función capaz de predecir el valor correspondiente a cualquier objeto de entrada válida después de haber visto una serie de ejemplos, los datos de entrenamiento. Para ello, tiene que generalizar a partir de los datos presentados a las situaciones no vistas previamente.

Los pasos que se deben considerar para conseguir la clasificación supervisada son los siguientes:

- Determinar el tipo de ejemplos de entrenamiento. Antes de hacer cualquier otra cosa, hay que decidir qué tipo de datos se va a utilizar para entrenar el modelo. Por ejemplo, podría ser un único carácter a mano, una palabra completa escrita a mano, o toda una línea de escritura a mano.
- Reunir un conjunto de entrenamiento. El conjunto de necesidades de formación a las características propias del uso del mundo real de la función. Por lo tanto, un conjunto de objetos de entrada que se recopilan y salidas correspondientes se recogen también, ya sea humana o de los expertos a partir de mediciones.
- Determinar la función de ingreso de la representación de la función aprendida. La precisión de la función aprendida depende en gran medida de cómo el objeto de entrada está representado. Normalmente, el objeto de entrada se transforma en un vector de características, que contiene una serie de características que son

descriptivos del objeto. El número de características no debe ser demasiado grande, a causa de la maldición de la dimensionalidad, pero debe ser lo suficientemente grande como para predecir con precisión la salida.

- Determinar la estructura de la función adecuada para resolver y el problema y la técnica de aprendizaje correspondiente. Por ejemplo, se podría optar por utilizar red neuronal artificial o un árbol de decisión.
- Ejecutar el algoritmo de aprendizaje en el conjunto de la formación obtenida. Parámetros del algoritmo de aprendizaje pueden ser ajustados mediante la optimización de rendimiento en un subconjunto de ellas (llamado conjunto de validación) del conjunto de entrenamiento, o por medio de la validación cruzada. Después del ajuste de parámetros y de aprendizaje, el desempeño del algoritmo se puede medir utilizando un conjunto de pruebas independiente del de entrenamiento.

### *Clasificación no supervisada*

Las clasificaciones no supervisadas son aquellas en las que el algoritmo clasificador no necesita de más información que la escena a clasificar y algunos parámetros que limiten el número de clases [4]. Estos mecanismos de clasificación basan su efecto en la búsqueda de clases con suficiente separabilidad espectral como para conseguir diferenciar unos elementos de otros.

El algoritmo más recurrente es el clustering. Este proceso consiste en la división de los datos en grupos de objetos similares. Para medir la similaridad entre objetos se suelen utilizar diferentes formas de distancia: distancia euclídea, de Manhatan, de Mahalanobis, etc. El representar los datos por una serie de clusters, conlleva la pérdida de detalles, pero consigue la simplificación de los mismos.

Desde un punto de vista práctico, el clustering juega un papel muy importante en aplicaciones de data mining, tales como exploración de datos científicos, recuperación de la información y minería de texto, aplicaciones sobre bases de datos espaciales (tales como GIS o datos procedentes de astronomía), aplicaciones Web , marketing, diagnóstico médico, análisis de ADN en biología computacional, y muchas otras. En el presente trabajo como ya hemos adelantado, nos centraremos en la política.

### *Clasificación paramétrica*

Se asume la forma del modelo y, a partir de los datos de entrenamiento, se hallan los valores adecuados para los parámetros del modelo [4].

Por ejemplo, un clasificador lineal asume que la clasificación puede realizarse mediante una combinación lineal de los valores de los atributos y emplea la combinación lineal que mejor se adapte al conjunto de casos de entrenamiento a la hora de clasificar nuevos casos.

En determinadas circunstancias, un clasificador cuadrático puede obtener mejores resultados que un clasificador lineal simple. Sin embargo, el ADC (Análisis Discriminante Cuadrático) requiere muchas más muestras de entrenamiento que el ADL (Análisis Discriminante Lineal) para obtener resultados similares ya que es más sensible al número de muestras requeridas.

#### *Clasificación no paramétrica*

No se conoce, o no se puede asumir, el conocimiento a priori de la estructura estadística de las clases.

Entre los distintos tipos de clasificadores no paramétricos se pueden destacar los que se basan en estimar las funciones de densidad de probabilidad, y los que estiman directamente la probabilidad a posteriori de la clase. Dentro de estos últimos, el más conocido es el clasificador basado en los k vecinos más cercanos [30]: si $k_i$ es el número de prototipos de la clase i entre los k más cercanos, la probabilidad a posteriori $P(\omega_i \mid x)$ se puede estimar como $\frac{k_i}{k}$. De esta manera, el clasificador asigna la muestra x a la clase más frecuente de entre sus k vecinos más cercanos, según una cierta medida de similitud o distancia.

#### *Clasificación simple*

Es un caso específico de clasificación binaria, donde solo se tiene una categoría para clasificar, es decir, pertenece a una categoría (con probabilidad *p*) o a la complementaria (con probabilidad *1-p*).

#### *Clasificación múltiple*

Tienen más de una categoría, por lo que se pueden solapar, al contrario que en la clasificación simple.

#### *Clasificación centrada en la categoría*

Dada una categoría se encuentran todos los documentos que estén clasificados dentro de dicha categoría.

#### *Clasificación centrada en el documento*

Dentro del documento se pueden encontrar todas las categorías en las que se pueden clasificar.

### 2.2.4. Algoritmos de clasificación automática de texto

Algoritmo: Según la RAE [3], es un conjunto ordenado y finito de operaciones que permite hallar la solución de un problema. Los algoritmos más frecuentes en el ámbito de la clasificación automática de textos son los mostrados a continuación:

#### *Algoritmo probabilístico*

Se basan en la teoría probabilística, en especial en el teorema de Bayes, el cual permite estimar la probabilidad de un suceso a partir de la probabilidad de que ocurra otro suceso, del cual depende el primero. El algoritmo más conocido, y también el más simple, es el denominado *Naïve Bayes* [17], que estima la probabilidad de que un documento pertenezca a una categoría.

Dicha pertenencia depende de la posesión de una serie de características, de cada una de las cuales se conoce la probabilidad de que aparezcan en los documentos que pertenecen a la categoría en cuestión. Naturalmente, dichas características son los términos que conforman

los documentos, y tanto su probabilidad de aparición en general, como la probabilidad de que aparezcan en los documentos de una determinada categoría, pueden obtenerse a partir de los documentos de entrenamiento; para ello se utilizan las frecuencias de aparición en la colección de entrenamiento.

Cuando las colecciones de aprendizaje son pequeñas, pueden producirse errores al estimar dichas probabilidades. Por ejemplo, cuando un determinado término no aparece nunca en esa colección de aprendizaje pero aparece en los documentos a categorizar. Esto implica la necesidad de aplicar técnicas de suavizado, a fin de evitar distorsiones en la obtención de las probabilidades.

Con dichas probabilidades, obtenidas de la colección de entrenamiento, podemos estimar la probabilidad de que un nuevo documento, dado que contiene un conjunto determinado de términos, pertenezca a cada una de las categorías. La más probable, obviamente, es a la que será asignado.

### *Algoritmo de Rocchio*

Cuando se expande de forma automática una consulta realizada por un usuario, esta puede ser realimentada utilizando aquellos documentos recuperados por la consulta inicial que el usuario señala como relevantes [18]. En estos casos, es preciso recalcular los pesos o importancia de los términos de la nueva consulta, o consulta expandida. El llamado algoritmo de Rocchio es un sistema de cálculo de dichos pesos, ampliamente utilizado. Desde un punto de vista práctico, su principal ventaja es que permite ajustar la importancia que se desea dar a los términos de los documentos relevantes de la consulta original, y también (en sentido negativo, obviamente) a los de los documentos que no se consideran relevantes.

Se construyen vectores que tratan de representar cada clase a partir de los documentos de entrenamiento. Para el vector de cada clase:

- Los documentos de entrenamiento de esa clase se usan como ejemplos positivos.
- Los documentos de entrenamiento de las demás clases se usan como ejemplos negativos.

El vector representativo de una clase se construye sumando los pesos de los términos de los ejemplos positivos. De él se restan los pesos de los términos de los ejemplos negativos. Aplicando coeficientes multiplicadores, es posible dar más o menos importancia a los ejemplos positivos o a los negativos. El resultado es un vector de términos con pesos como el utilizado en el modelo vectorial. Para clasificar un nuevo documento, no hay más que estimar la similitud entre el vector de ese documento y los vectores de cada una de las clases.

### *Algoritmo de vecino más próximo (k-NN)*

El método k-nn [21] es un método de aprendizaje inductivo supervisado que sirve para estimar la función de densidad $F(x / C_j)$ de las predictoras x por cada clase $C_j$ . Este es un método de clasificación no paramétrico, que estima el valor de la función de densidad de probabilidad o directamente la probabilidad a posteriori de que un elemento x pertenezca a la clase $C_j$ a partir de la información proporcionada por el conjunto de prototipos. En el proceso

de aprendizaje no se hace ninguna suposición acerca de la distribución de las variables predictoras.

En el reconocimiento de patrones, el algoritmo k-nn es usado como método de clasificación de objetos (elementos) basado en un entrenamiento mediante ejemplos cercanos en el espacio de los elementos. k-nn es un tipo de "Lazy Learning", donde la función se aproxima solo localmente y todo el cómputo es diferido a la clasificación.

Los ejemplos de entrenamiento son vectores en un espacio característico multidimensional. Cada ejemplo está descrito en términos de *p* atributos considerando *q* clases para la clasificación. Los valores de los atributos del *i-esimo* ejemplo (donde $1 \leq i \leq n$ ) se representan por el vector *p* dimensional

$$x_i = (x_{1i}, x_{2i}, \dots, x_{pi}) \in X$$

El espacio es particionado en regiones por localizaciones y etiquetas de los ejemplos de entrenamiento. Un punto en el espacio es asignado a la clase C si esta es la clase más frecuente entre los k ejemplos de entrenamiento más cercano. Generalmente se usa la Distancia euclídea.

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{p} (x_{ir} - x_{jr})^2}$$

La fase de entrenamiento del algoritmo consiste en almacenar los vectores característicos y las etiquetas de las clases de los ejemplos de entrenamiento. En la fase de clasificación, la evaluación del ejemplo (del que no se conoce su clase) es representada por un vector en el espacio característico. Se calcula la distancia entre los vectores almacenados y el nuevo vector, y se seleccionan los k ejemplos más cercanos. El nuevo ejemplo es clasificado con la clase que más se repite en los vectores seleccionados.

Este método supone que los vecinos más cercanos nos dan la mejor clasificación y esto se hace utilizando todos los atributos; el problema de dicha suposición es que es posible que se tengan muchos atributos irrelevantes que dominen sobre la clasificación: dos atributos relevantes perderían peso entre otros veinte irrelevantes.

Para corregir el posible sesgo se puede asignar un peso a las distancias de cada atributo, dándole así mayor importancia a los atributos más relevantes. Otra posibilidad consiste en tratar de determinar o ajustar los pesos con ejemplos conocidos de entrenamiento. Finalmente, antes de asignar pesos es recomendable identificar y eliminar los atributos que se consideran irrelevantes.

En la Figura 6 se desea clasificar el círculo verde. Para k = 3 este es clasificado con la clase triángulo, ya que hay solo un cuadrado y 2 triángulos, dentro del círculo que los contiene. Si k = 5 este es clasificado con la clase cuadrado, ya que hay 2 triángulos y 3 cuadrados, dentro del círculo externo.

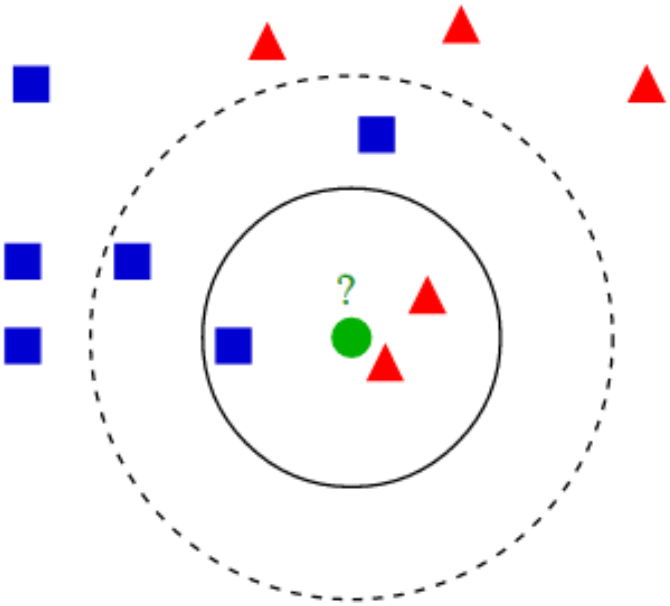



Figura 6: Ejemplo de clasificación mediante k-NN.

La mejor elección de k depende fundamentalmente de los datos; generalmente, valores grandes de k reducen el efecto de ruido en la clasificación, pero crean límites entre clases parecidas. Un buen k puede ser seleccionado mediante una optimización de uso. El caso especial en que la clase es predicha para ser la clase más cercana al ejemplo de entrenamiento (cuando k = 1) es llamada Nearest Neighbor Algorithm, o traducido al castellano, Algoritmo del vecino más cercano.

La exactitud de este algoritmo puede ser severamente degradada por la presencia de ruido o características irrelevantes, o si las escalas de características no son consistentes con lo que uno considera importante. Muchas investigaciones y esfuerzos fueron puestos en la selección y crecimiento de características para mejorar las clasificaciones. Particularmente una aproximación en el uso de algoritmos que evolucionan para optimizar características de escalabilidad. Otra aproximación consiste en escalar características por la información mutua de los datos de entrenamiento con las clases de entrenamiento.

### 2.2.5. Métodos de aprendizaje

#### *Árboles de decisión*

Los árboles de decisión [5] (también llamados de clasificación o de identificación) constituyen una aproximación radicalmente distinta a todas las estudiadas hasta el momento. Es uno de los métodos de aprendizaje inductivo supervisado no paramétrico más utilizado. Como forma de representación del conocimiento, los árboles de clasificación destacan por su sencillez. A pesar de que carecen de la expresividad de las redes semánticas o de la lógica de primer orden, su dominio de aplicación no está restringido a un ámbito concreto sino que

pueden utilizarse en diversas áreas: diagnóstico médico, juegos, predicción meteorológica, control de calidad, etc.

Un árbol de clasificación es una forma de representar el conocimiento obtenido en el proceso de aprendizaje inductivo. Puede verse como la estructura resultante de la partición recursiva del espacio de representación a partir del conjunto (numeroso) de prototipos. Esta partición recursiva se traduce en una organización jerárquica del espacio de representación que puede modelarse mediante una estructura de tipo árbol. Cada nodo interior contiene una pregunta sobre un atributo concreto (con un hijo por cada posible respuesta) y cada nodo hoja se refiere a una decisión (clasificación).

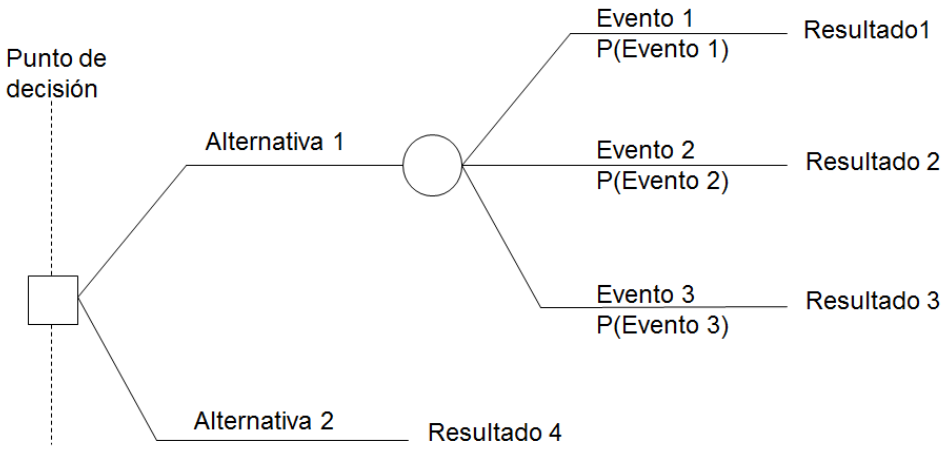



Figura 7: Ejemplo de árbol de decisión.

La clasificación de patrones se realiza en base a una serie de preguntas sobre los valores de sus atributos, empezado por el nodo raíz y siguiendo el camino determinado por las respuestas a las preguntas de los nodos internos, hasta llegar a un nodo hoja. La etiqueta asignada a esta hoja es la que se asignará al patrón a clasificar.

La metodología a seguir para la construcción de un árbol de decisión, puede resumirse en dos pasos, y se esquematiza en la Figura 8:

- Aprendizaje: Consiste en la construcción del árbol a partir de un conjunto de prototipos, S. Constituye la fase más compleja y la que determina el resultado final.
- Clasificación: Consiste en el etiquetado de un patrón, X, independiente del conjunto de aprendizaje. Se trata de responder a las preguntas asociadas a los nodos interiores utilizando los valores de los atributos del patrón X. Este proceso se repite desde el nodo raíz hasta alcanzar una hoja, siguiendo el camino impuesto por el resultado de cada evaluación.

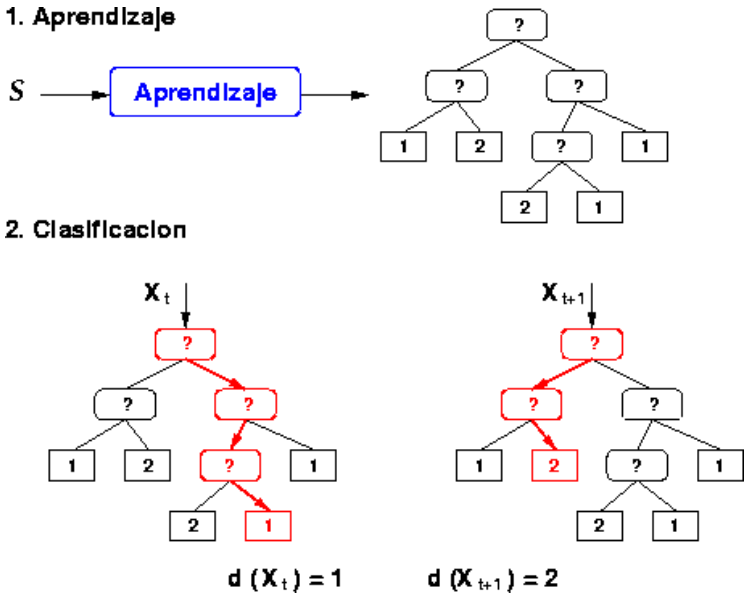



Figura 8: Aprendizaje y clasificación mediante un árbol de decisión.

### *Máquinas de vectores de soporte*

Las máquinas de vectores de soporte [19] (SVM, del inglés *Support Vector Machines*) tienen su origen en los trabajos sobre la teoría del aprendizaje estadístico y fueron introducidas en los años 90 por Vapnik y sus colaboradores. Aunque originalmente las SVMs fueron pensadas para resolver problemas de clasificación binaria, actualmente se utilizan para resolver otros tipos de problemas (regresión, agrupamiento, multiclasificación). También son diversos los campos en los que han sido utilizadas con éxito, tales como visión artificial, reconocimiento de caracteres, categorización de texto e hipertexto, clasificación de proteínas, procesamiento de lenguaje natural, análisis de series temporales. De hecho, desde su introducción, han ido ganando un merecido reconocimiento gracias a sus sólidos fundamentos teóricos.

Dentro de la tarea de clasificación, las SVMs pertenecen a la categoría de los clasificadores lineales, puesto que inducen separadores lineales o hiperplanos, ya sea en el espacio original de los ejemplos de entrada, si estos son separables o cuasi-separables (ruido), o en un espacio transformado (espacio de características), si los ejemplos no son separables linealmente en el espacio original.

Mientras la mayoría de los métodos de aprendizaje se centran en minimizar los errores cometidos por el modelo generado a partir de los ejemplos de entrenamiento (error empírico), el sesgo inductivo asociado a las SVMs radica en la minimización del denominado riesgo estructural. La idea es seleccionar un hiperplano de separación que equidista de los ejemplos más cercanos de cada clase para, de esta forma, conseguir lo que se denomina un margen máximo a cada lado del hiperplano. Además, a la hora de definir el hiperplano, solo se consideran ejemplos de entrenamiento de cada clase que caen justo en la frontera de dichos márgenes. Estos ejemplos reciben el nombre de vectores soporte. Desde un punto de vista práctico, el hiperplano separador de margen máximo ha demostrado tener una buena capacidad de generalización, evitando en gran medida el problema de sobreajuste a los ejemplos de entrenamiento.

Desde un punto de vista algorítmico, el problema de optimización del margen geométrico representa un problema de optimización cuadrático con restricciones lineales que puede ser resuelto mediante técnicas estándar de programación cuadrática. La propiedad de convexidad exigida para su resolución garantiza una solución única, en contraste con la no unicidad de la solución producida por una red neuronal artificial entrenada con un mismo conjunto de ejemplos.

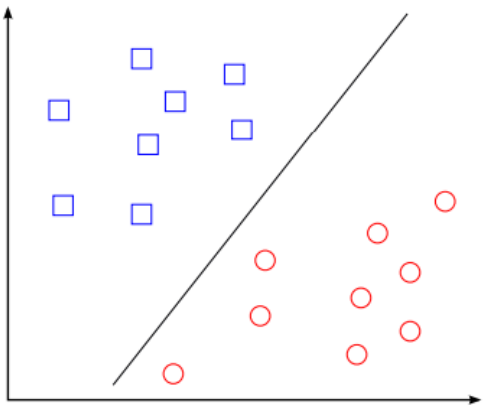


Figura 9: Ejemplo de hiperplano de separación.

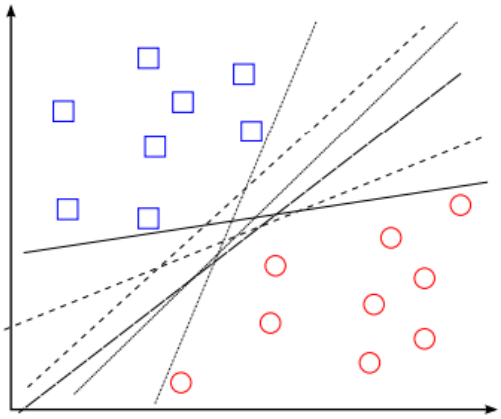


Figura 10: Ejemplo de hiperplano de separación, de entre los infinitos posibles.

### *Redes neuronales*

Los primeros modelos de redes neuronales artificiales (RNA) [11] datan de 1943 por los neurólogos Warren McCulloch y Walter Pitts. Años más tarde, en 1949, Donald Hebb desarrolló sus ideas sobre el aprendizaje neuronal, quedando reflejado en la "regla de Hebb". En 1958, Rosenblatt desarrolló el perceptrón simple, y en 1960, Widrow y Hoff desarrollaron el ADALINE, que fue la primera aplicación industrial real.

En los años siguientes, se redujo la investigación, debido a la falta de modelos de aprendizaje y el estudio de Minsky y Papert sobre las limitaciones del perceptrón. Sin embargo, en los años 80, volvieron a resurgir las RNA gracias al desarrollo de la red de Hopfield, y en especial, al algoritmo de aprendizaje de retropropagación (BackPropagation) ideado por Rumelhart y McClelland en 1986 que fue aplicado en el desarrollo de los perceptrones multicapa.

A pesar de su nombre, las redes neuronales no tienen un concepto demasiado complicado detrás de ellas. El nombre viene de la idea de imitar el funcionamiento de las redes neuronales de los organismos vivos: un conjunto de neuronas conectadas entre sí y que trabajan en conjunto, sin que haya una tarea concreta para cada una. Con la experiencia, las neuronas van creando y reforzando ciertas conexiones para "aprender" algo que se queda fijo en el tejido.

Ahora bien, el enfoque biológico no ha sido especialmente útil: las redes neuronales han ido moviéndose para tener un foco en matemáticas y estadística. Se basan en una idea sencilla: dados unos parámetros hay una forma de combinarlos para predecir un cierto resultado. Por ejemplo, sabiendo los píxeles de una imagen habrá una forma de saber qué

número hay escrito, o conociendo la carga de servidores de un Centro de Procesamiento de Datos (CPD), su temperatura y demás existirá una manera de saber cuánto van a consumir.

Las redes neuronales son un modelo para encontrar esa combinación de parámetros y aplicarla al mismo tiempo. En el lenguaje propio, encontrar la combinación que mejor se ajusta es "entrenar" la red neuronal. Una red ya entrenada se puede usar luego para hacer predicciones o clasificaciones, es decir, para "aplicar" la combinación.

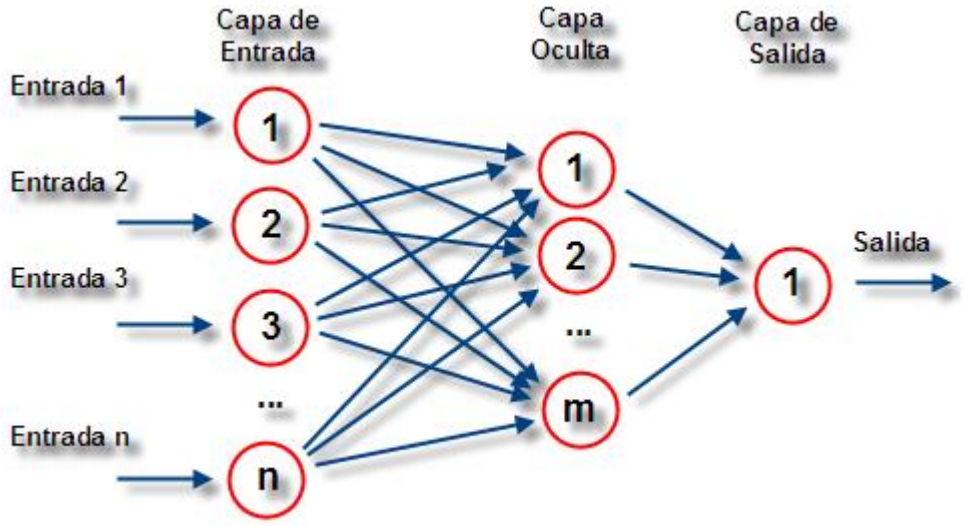



**Figura 11: Red neuronal artificial perceptrón con n neuronas de entrada, m neuronas en su capa oculta y una neurona de escape.**

## 2.3. Clasificador basado en reglas

Hoy en día, la mayoría de la información generada se almacena para su posterior consulta y/o procesamiento. Igualmente, en las redes de computadoras se pueden analizar los datos proporcionados por el protocolo TCP/IP para detectar intrusos o comportamientos anómalos. La cantidad de información almacenada por los sistemas actuales es muy grande para ser analizada manualmente.

La Minería de Datos ofrece herramientas para descubrir información implícita en estos grandes conjuntos de datos. Una importante técnica de la Minería de Datos es el descubrimiento o minado de reglas de asociación que permite descubrir relaciones interesantes, llamadas asociaciones, en grandes conjuntos de datos.

Desde finales de los 90 se comenzó a investigar el poder discriminativo de las reglas de asociación y estas se utilizaron para construir clasificadores de alta eficacia. Estos clasificadores se construyen a partir de un conjunto especial de reglas denominadas Reglas de Asociación de Clase.

La clasificación con reglas de asociación se ha aplicado en diferentes tareas como son: la reducción de fallas en las telecomunicaciones, la detección de redundancia en exámenes médicos [15] o la clasificación de textos [16].

Aprovechando el poder discriminativo de las reglas de asociación, se comienzan a integrar las técnicas de *Classification Rule Mining* (CRM) y *Association Rule Mining* (ARM) [27]. La integración de ambas técnicas consiste en minar un subconjunto especial de reglas de asociación denominadas reglas de asociación de clase y utilizar este subconjunto para construir clasificadores. Los clasificadores desarrollados desde entonces se dividen en dos grupos: Los clasificadores de dos etapas y los clasificadores integrados.

**Clasificadores de dos etapas**: Estos clasificadores, en la primera etapa calculan todas las reglas de asociación de clase. En la segunda etapa se determina un subconjunto más pequeño de reglas de asociación de clase que cubra al conjunto de entrenamiento y con este se construye el clasificador [27].

**Clasificadores integrados**: Los clasificadores integrados utilizan diferentes estrategias para generar directamente el conjunto de reglas de asociación de clase, construyendo el clasificador en una sola etapa.

## 2.4. Redes sociales

El concepto de red social ha adquirido una importancia notable en los últimos años. Se ha convertido en una expresión del lenguaje común que asociamos a nombres como Facebook o Twitter. Pero su significado es mucho más amplio y complejo. Las redes sociales son, desde hace décadas, objeto de estudio de numerosas disciplinas. Alrededor de ellas se han generado teorías de diverso tipo que tratan de explicar su funcionamiento y han servido, además, de base para su desarrollo virtual. Con la llegada de la Web 2.0, las redes sociales en Internet ocupan un lugar relevante en el campo de las relaciones personales y son, asimismo, paradigma de las posibilidades que nos ofrece esta nueva forma de usar y entender Internet. Vamos a definir las redes sociales teniendo en cuenta todos estos matices con el fin de entenderlas mejor como fenómeno y herramienta.


Figura 12: Redes sociales.

En sentido amplio, una red social es una estructura social formada por personas o entidades conectadas y unidas entre sí por algún tipo de relación o interés común. El término se atribuye a los antropólogos británicos Alfred Radcliffe-Brown y John Barnes [33]. Las redes sociales son parte de nuestra vida, son la forma en la que se estructuran las relaciones personales, estamos conectados mucho antes de tener conexión a Internet. En antropología y sociología, las redes sociales han sido materia de estudio en diferentes campos, desde el análisis de las relaciones de parentesco en grupos pequeños hasta las nuevas investigaciones sobre diásporas de inmigrantes en entornos multisituados. Pero el análisis de las redes sociales también ha sido llevado a cabo por otras especialidades que no pertenecen a las ciencias sociales. Por ejemplo, en matemáticas y ciencias de la computación, la teoría de grafos representa las redes sociales mediante nodos conectados por aristas, donde los nodos serían los individuos y las aristas las relaciones que les unen. Todo ello conforma un grafo, una estructura de datos que permite describir las propiedades de una red social. A través de esta teoría, se pueden analizar las redes sociales existentes entre los empleados de una empresa y, de igual manera, entre los amigos de Facebook.

### 2.4.1. Historia y evolución de las redes sociales

Trazar la historia de las redes sociales no es una tarea fácil, su origen es difuso y su evolución acelerada. No existe consenso sobre cuál fue la primera red social, y podemos encontrar diferentes puntos de vista al respecto. Por otro lado, la existencia de muchas plataformas se cuenta en tiempos muy cortos, bien sabido es que hay servicios de los que hablamos hoy que quizá mañana no existan, y otros nuevos aparecerán dejando obsoleto, en poco tiempo, cualquier panorama que queramos mostrar de ellos. Su historia se escribe a cada minuto en cientos de lugares del mundo. Lo que parece estar claro es que los inicios se remontan mucho más allá de lo que podríamos pensar en un primer momento, puesto que los primeros intentos de comunicación a través de Internet ya establecen redes, y son la semilla que dará lugar a lo que más tarde serán los servicios de redes sociales que conocemos actualmente, con creación de un perfil y lista de contactos. Por todo ello, vamos a plantear su historia [33] contextualizada mediante una cronología de los hechos más relevantes del fenómeno que suponen las redes sociales basadas en Internet.

- ***1971***. Se envía el primer e-mail entre dos ordenadores situados uno al lado del otro.
- ***1978***. Ward Christensen y Randy Suess crean el **BBS** (Bulletin Board Systems) para informar a sus amigos sobre reuniones, publicar noticias y compartir información.
- ***1994***. Se lanza **GeoCities**, un servicio que permite a los usuarios crear sus propios sitios web y alojarlos en determinados lugares según su contenido.
- ***1995***. La Web alcanza el millón de sitios web, y **The Globe** ofrece a los usuarios la posibilidad de personalizar sus experiencias on-line, mediante la publicación de su propio contenido y conectando con otros individuos de intereses similares. En este mismo año, Randy Conrads crea **Classmates**, una red social para contactar con antiguos compañeros de estudios. Classmates es para muchos el primer servicio de red social, principalmente, porque se ve en ella el germen de Facebook y otras redes sociales que nacieron, posteriormente, como punto de encuentro para alumnos y ex-alumnos.
- ***1997***. Lanzamiento de **AOL Instant Messenger**, que ofrece a los usuarios el chat, al tiempo que comienza el **blogging** y se lanza **Google**. También se inaugura **Sixdegrees**, red social que permite la creación de perfiles personales y listado de amigos, algunos establecen con ella el inicio de las redes sociales por reflejar mejor sus funciones características. Solo durará hasta el año 2000.
- ***1998***. Nace **Friends Reunited**, una red social británica similar a Classmates. Asimismo, se realiza el lanzamiento de **Blogger**.
- ***2000***. Estalla la **"Burbuja de Internet"**. En este año se llega a la cifra de setenta millones de ordenadores conectados a la Red.
- ***2002***. Se lanza el portal **Friendster**, que alcanza los tres millones de usuarios en solo tres meses.
- ***2003***. Nacen **MySpace, LinkedIn** y **Facebook**, aunque la fecha de esta última no está clara puesto que llevaba gestándose varios años. Creada por el conocido Mark Zuckerberg, Facebook se concibe inicialmente como plataforma para conectar a los estudiantes de la Universidad de Harvard. A partir de este momento nacen muchas otras redes sociales como **Hi5** y **Netlog**, entre otras.

- ***2004***. Se lanzan **Digg**, como portal de noticias sociales; **Bebo**, con el acrónimo de "Blog Early, Blog Often"; y **Orkut**, gestionada por Google.
- ***2005.*** **YouTube** comienza como servicio de alojamiento de vídeos, y **MySpace** se convierte en la red social más importante de Estados Unidos.
- ***2006***. Se inaugura la red social de microblogging **Twitter**. **Google** cuenta con 400 millones de búsquedas por día, y **Facebook** sigue recibiendo ofertas multimillonarias para comprar su empresa. En España se lanza **Tuenti**, una red social enfocada al público más joven. Este mismo año, también comienza su actividad **Badoo**.
- ***2008***. **Facebook** se convierte en la red social más utilizada del mundo con más de 200 millones de usuarios, adelantando a **MySpace**. Nace **Tumblr** como red social de microblogging para competir con Twitter.
- ***2009***. **Facebook** alcanza los 400 millones de miembros, y **MySpace** retrocede hasta los 57 millones. El éxito de Facebook es imparable.
- ***2010***. Google lanza **Google Buzz**, su propia red social integrada con Gmail, en su primera semana sus usuarios publicaron nueve millones de entradas. También se inaugura otra nueva red social, **Pinterest**. Los usuarios de **Internet** en este año se estiman en 1,97 billones, casi el 30% de la población mundial. Las cifras son asombrosas: **Tumblr** cuenta con dos millones de publicaciones al día; **Facebook** crece hasta los 550 millones de usuarios: **Twitter** computa diariamente 65 millones de tweets, mensajes o publicaciones de texto breve; **LinkedIn** llega a los 90 millones de usuarios profesionales, y **YouTube** recibe dos billones de visitas diarias.

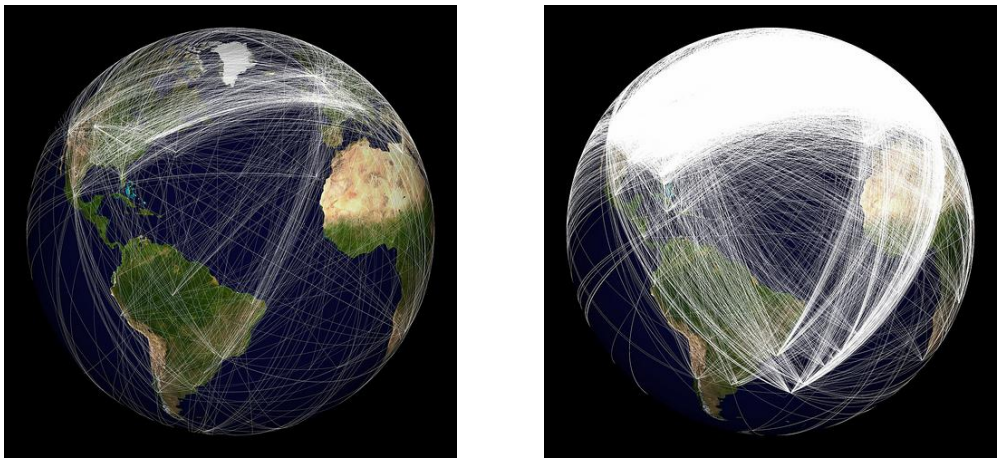


**Figura 13. Conexiones de las redes sociales en el mundo en el año 2009, izquierda, y en el año 2010, derecha.**

- ***2011***. **MySpace** y **Bebo** se rediseñan para competir con **Facebook** y **Twitter**. **LinkedIn** se convierte en la segunda red social más popular en Estados Unidos con 33,9 millones de visitas al mes. En este año se lanza **Google+**, otra nueva apuesta de Google por las redes sociales. La recién creada **Pinterest** alcanza los diez millones de visitantes mensuales. **Twitter** multiplica sus cifras rápidamente y en solo un año aumenta los tweets recibidos hasta los 33 billones.

- ***2012***. **Facebook** supera los 800 millones de usuarios, **Twitter** cuenta con 200 millones, y **Google+** registra 62 millones. La red española **Tuenti** alcanzó en febrero de este año los 13 millones de usuarios.
- **2013**. Se lanza **Vine,** aplicación comprada por **Twitter** que permite crear y publicar vídeos cortos, de una duración máxima de seis segundos, en forma de loop (reproducción continua).
- **2015**. A comienzos de este año, el crecimiento de las redes sociales se antoja imparable. **Facebook** supera la cifra de 1,2 billones de usuarios activos. **YouTube** alcanza el billón, mientras que **Google+** adelanta a **Twitter**, con 540 millones de usuarios activos frente a los más de 250 millones de esta. Este crecimiento se debe en gran medida a que Google "fuerza" a todos los usuarios a generar una cuenta para su enorme variedad de productos hegemónicos en sus áreas.

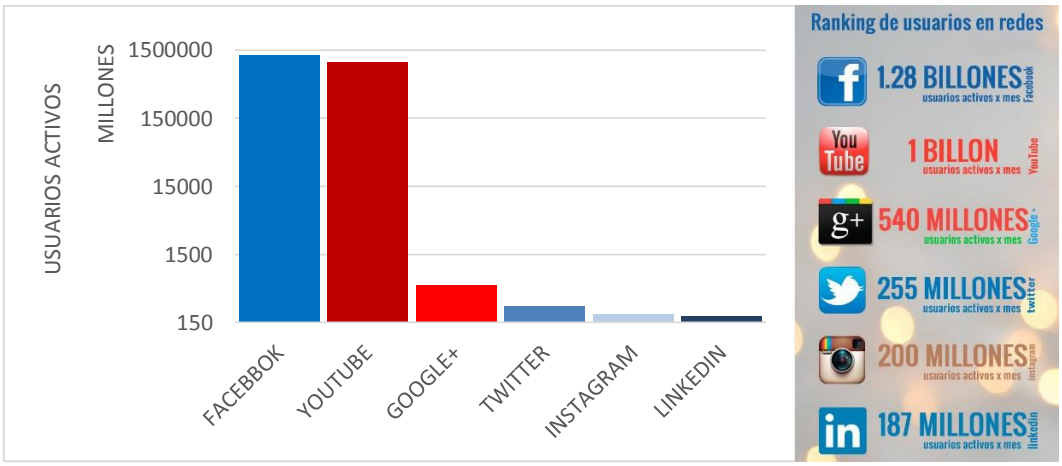



Figura 14. Usuarios activos en las redes sociales enero-2015 [12].

En España, las redes sociales más usadas no difieren demasiado si comparamos los datos anteriores. **Facebook** es la red social preferida de los españoles seguida de **YouTube**. Sin embargo, la tercera red más usada es **Twitter**, seguida de cerca por **Google+.**


Figura 15. Ranking de uso de las RRSS en España [26].

### 2.4.2. Facebook

Facebook es una red social gratuita. Inicialmente se creó para estudiantes de la Universidad de Harvard en Estados Unidos. Era un instrumento que permitía que los alumnos tuvieran contacto entre ellos, intercambiar notas sobre sus curso e incluso organizar todo tipo de reuniones estudiantiles. En septiembre del 2006 se abrió a toda persona que tuviera un email o correo electrónico lo que elevó a 140 millones de usuarios ese año.

Fue creado por Mark Zuckerberg [6] y fundada por él y Eduardo Saverin, Dustin Moskovitz y Chris Hughes. A partir del año 2007 lanza versiones en francés, alemán y español que fueron traducidas por usuarios no remunerados, ya que su finalidad era impulsar esta red fuera de Estados Unidos, porque la mayoría de sus usuarios estaban en ese país, Reino Unido y Canadá. Es la red social más popular y tiene cada vez más usuarios móviles. Los países con más usuarios son Brasil, India, Indonesia, México y Estados Unidos.


**Figura 16. Mark Zuckerberg, creador y fundador de Facebook.**

Facebook es gratuito para los usuarios y genera ingresos por la publicidad expuesta, incluyendo los banners y los grupos patrocinados. Los usuarios pueden registrarse a través de su correo electrónico y pueden hacerlo como celebridades, músicos o grupos de música, negocios o empresas, o personas individuales. Pueden crear perfiles que contienen fotos, listas de intereses personales e información privada o no, y pueden realizar un intercambio de mensajes privados y públicos entre sí y en los grupos de amigos. La visualización de los datos detallados de los miembros está restringida a los miembros de la misma red, a los amigos confirmados, o puede ser libre para cualquier persona.

### 2.4.3. YouTube

Desde mayo de 2005, miles de millones de usuarios encuentran, ven y comparten vídeos originales en YouTube. YouTube se ha convertido en un foro donde los usuarios pueden

interactuar, obtener información e inspirar a otras personas de todo el mundo, y sirve de plataforma de distribución para creadores de contenido original y para anunciantes grandes y pequeños.

Fue creado por tres antiguos empleados de PayPal [25], Chad Hurley, Steve Chen y Jawed Karim en febrero de 2005. En octubre de 2006, fue adquirido por Google Inc. a cambio de 1650 millones de dólares y ahora opera como una de sus filiales. Actualmente es el sitio web de su tipo más utilizado en Internet.

**Figura 17. Chad Hurley, co-creador de YouTube.**

YouTube es un reproductor en línea basado en el estándar HTML5. Este estándar se incorporó poco después de que la W3C lo presentara y que es soportado por los navegadores web más difundidos. Los enlaces a vídeos de YouTube pueden ser también insertados en blogs y sitios electrónicos personales usando la API o incrustando cierto código HTML.

### 2.4.4. Twitter

Twitter es una de las redes sociales de mayor crecimiento, basada en el concepto de "microblogging" [13], que permite a los usuarios postear mensajes de una longitud reducida en número de caracteres. A través de las APIs de Twitter cualquiera puede crear aplicaciones que comuniquen con el servicio de la mencionada red social.

El 21 de marzo del 2006, Jack Dorsey lanzó al mundo el primer tweet. Una primera y simple invitación (“inviting coworkers“) para entrar en lo que por aquel entonces se trataba de un proyecto en ciernes.


Figura 18. Jack Dorsey.

La red permite enviar mensajes de longitud corta, con un máximo de 140 caracteres, llamados tweets, que se muestran en la página principal del usuario. Los usuarios pueden suscribirse a los tweets de otros usuarios. A los usuarios abonados se les llama "seguidores", o “followers".

Por defecto, los mensajes son públicos, pudiendo difundirse privadamente mostrándolos únicamente a unos seguidores determinados. Los usuarios pueden twittear desde la web del servicio, con aplicaciones oficiales externas (como para teléfonos inteligentes), o mediante el Servicio de mensajes cortos (SMS) disponible en ciertos países.

A finales del primer semestre de este año, Twitter alcanzó los 316 millones de usuarios activos, enviando aproximadamente 500 millones de tweets diarios.


**Figura 19. Twitter en números [13].**

### 2.4.5. Política y Twitter

Desde hace años, se antoja evidente el poder que está adquiriendo Internet como herramienta de comunicación política y electoral. Ya no se concibe una campaña sin el uso planificado y sistemático de este medio y ningún candidato quiere quedarse al margen de las nuevas posibilidades comunicativas que brinda la Red.

Y de entre todas las posibilidades que se plantean, sin duda la revolución de las redes sociales está causando un gran impacto en el panorama político actual. De todas estas redes, Twitter es la que actualmente más relevancia ha adquirido entre la clase política y periodística. En palabras de Piscitelli, se ha convertido "en uno de los mecanismos de comunicación más poderosos de la historia" [32]. Y según Orihuela, "ha cambiado la red y ha completado el giro social que iniciaron los blogs a finales de los años noventa" [31].

Una nueva escena política se ha planteado en los últimos meses en España, donde se han sumado los nuevos partidos políticos como Ciudadanos o Podemos, rompiendo así el llamado Bipartidismo político que regentaban Partido Popular (PP), y Partido Socialista (PSOE). Esta confrontación por la lucha del voto se ve reflejada constantemente en las redes sociales. En época de elecciones electorales, tanto los partidos políticos como sus máximos representantes multiplican sus comentarios en las redes sociales, con el afán de captar la atención de un gran número de ciudadanos y de atraer futuros votantes.

Twitter es probablemente una de las redes sociales donde más se hace notar esta lucha. Tanto es así que Twitter ha sido protagonista directo de cambios en grupos de gobierno, como sucedió en el Ayuntamiento de la Comunidad de Madrid, cuando Guillermo Zapata dimitió como edil de Cultura por el "dolor generado" por sus tweets [23].

Por lo tanto, parece acertado enfocar el trabajo para la clasificación automática de textos para el seguimiento de campañas electorales en una red social con tanta repercusión mediática y tanto seguimiento ciudadano como puede ser Twitter.

### 2.4.6. Herramientas de monitorización y análisis

Existen varias herramientas para la gestión, monitorización y análisis de datos en Twitter. Los usuarios las utilizan para generar contenido de calidad, de forma que son los propios usuarios que pasan de ser receptores pasivos a informarse de forma activa, sin interrupciones publicitarias de los medios.

La respuesta a estos cambios por parte de los anunciantes son nuevas formas de entender el marketing en la que los consumidores son los que encuentran las marcas e interactúan con ellas de forma consentida sin producirse interrupciones no deseadas.

Algunas de las herramientas de gestión [1] son:

- **Hootsuite**: esta herramienta permite gestionar una o varias cuentas de Twitter y de otras redes sociales como Google+ o Facebook. Su estructura de columnas permite monitorizar menciones, mensajes privados, el timeline, búsquedas relevantes para un negocio, personas hablando de marcas y listas de Twitter.
- **TweetDeck**: se trata de una herramienta de escritorio similar a Hootsuite que permite realizar la gestión completa. Destaca por su usabilidad.
- **BufferApp**: se trata de una herramienta con una interfaz sencilla cuyo principal objeto es programar actualizaciones a las horas que decidas casi sin esfuerzo y medir el resultado de cada tweet (alcance, repuestas, favoritos y número de retweets).

Herramientas de monitorización [1]:

- **Mention.net**: se trata de una herramienta de monitorización en Internet con la que se puede estar al tanto de todas las menciones que se produzcan de una marca, empresa, nombre en redes sociales, la web o blogs. Su precio varía según el volumen de menciones que se generen de forma mensual.
- **SocialMention**: similar a Mention.net permite monitorizar toda la actividad en internet de forma sencilla.
- **Google Alerts**: este servicio de Google permite generar notificaciones cuando el motor de búsqueda indexa contenido que coincide con las palabras clave y criterios configurados. Lo cierto es que desde 2012 su eficacia está en seria duda y su uso es cada vez menor.

En cuanto a herramientas de análisis destacan [1]:

- **Topsy**: Topsy es un motor de búsqueda que permite conocer toda la actividad que se produce en Twitter.

- **Tweriod**: Esta herramienta, se encarga de analizar las franjas horarias en las que los seguidores están más activos para poder publicar en los momentos en el que mayor visibilidad tendrán.
- **TweetReach**: es una excelente herramienta para medir el número de impresiones y el alcance de los hashtags.
- **Klout**: aunque polémico por los criterios que utiliza, Klout es un muy sencillo indicador de la autoridad de cada cuenta de Twitter. Con una escala de 0 (mínimo) a 100 nos indica cómo de influyente es cada usuario.
- **FollowerWonk**: se trata de un motor de búsqueda dentro de Twitter. Mediante búsquedas basadas en palabras clave y/o localización, esta herramienta muestra las cuentas que se ajustan a esos criterios para poder filtrar por número de seguidores, tweets y número de cuentas a las que siguen y poder conectar con las personas.
- **Google Analytics**: como cualquier rama del marketing online, es vital medir si Twitter está llevando tráfico y conversiones al sitio web. Google Analytics permite hacer ese seguimiento en fuentes de tráfico.

# 3. Marco regulador

El problema legislativo que puede surgir con este Trabajo Fin de Grado, puede encuadrarse dentro del ámbito de la gestión de la información, ya que al tratarse de un clasificador de tweets, puede que se incumpla la legislación vigente de la protección de datos, ya que los tweets contienen información personal, de carácter público o privado.

En las fases de recopilación de datos, implementación y validación del clasificador, se ha tratado con tweets. Es, por tanto, en este aspecto donde analizaremos cuidadosamente la legislación vigente con la finalidad de no infringir ninguna ley.

Consultamos la legislación vigente en la Agencia Española de Protección de Datos [2]. En el artículo 3 del título primero, citamos textualmente, de La ley de protección de datos [9]:

*"A los efectos de la presente Ley Orgánica se entenderá por:*
*a) Datos de carácter personal: cualquier información concerniente a personas físicas identificadas o identificables."*

Aunque nosotros solo veamos 140 caracteres, detrás de cada tweet hay muchísima más información técnica de la que nos podemos imaginar (metadatos). A modo de curiosidad adjuntamos una imagen en el anexo, que hace referencia a la información que queda plasmada en un tweet [Anexo].

Cabe destacar que los metadatos que aluden a la posible identificación del usuario son: *name, user, description, image, geo, coordinates, place y text.*

Con los metadatos de geo, coordinates y place; se puede averiguar desde donde se envió el tweet. Esto nos lleva a pensar que con este tipo de metadatos podríamos identificar al usuario, ya que podríamos averiguar sus rutinas o lugar de estudios/trabajo. El usuario es libre de ocultar, o no, este tipo de información.

Los metadatos name, user e image, son relativos, ya que depende del usuario la información que quiera transmitir. Hay cuentas de usuarios que sí facilitan su nombre real y cuentas en las que se utilizan sobrenombres. Con las fotos pasa algo similar, ya que hay cuentas en las que la foto es del usuario (usuario con amigos o familiares), y hay cuentas en las que la foto no tiene nada que ver con su usuario. En cualquier caso son datos que el usuario aporta y es consciente que la información es visible a todo aquel que accede a su perfil.

Por otro lado, en el artículo 4 de título segundo podemos leer lo siguiente:

*"2. Los datos de carácter personal objeto de tratamiento no podrán usarse para finalidades incompatibles con aquellas para las que los datos hubieran sido recogidos. No se considerará incompatible el tratamiento posterior de estos con fines históricos, estadísticos o científicos."*

Nos indica claramente que los análisis de los textos quedan protegidos si tienen finalidad estadísticos o científicos, como consecuencia también quedan protegidos los resultados de este estudio.

Y si no quedaba claro si estamos o no infringiendo la ley, encontramos en el artículo 6 del título 2 apartado 2 lo siguiente:

*"2. No será preciso el consentimiento cuando los datos de carácter personal se recojan para el ejercicio de las funciones propias de las Administraciones públicas en el ámbito de sus competencias; cuando se refieran a las partes de un contrato o precontrato de una relación negocial, laboral o administrativa y sean necesarios para su mantenimiento o cumplimiento; cuando el tratamiento de los datos tenga por finalidad proteger un interés vital del interesado en los términos del artículo 7, apartado 6, de la presente Ley, o cuando los datos figuren en fuentes accesibles al público y su tratamiento sea necesario para la satisfacción del interés legítimo perseguido por el responsable del fichero o por el del tercero a quien se comuniquen los datos, siempre que no se vulneren los derechos y libertades fundamentales del interesado."*

En conclusión, Twitter es una fuente de información en la que por defecto, los mensajes son públicos, pudiendo difundirse por privado mostrándolos únicamente a unos seguidores determinados, y por lo tanto el metadato text, que es el que utilizamos para poner a prueba nuestro clasificador, es de carácter público. Deducimos por lo tanto que respetamos la legislación vigente en el país.

# 4. Captura de datos

## 4.1. Generalidades de la API de Twitter

La interfaz de programación de aplicaciones, abreviada como API [20] (del inglés: Application Programming Interface), es el conjunto de subrutinas, funciones y procedimientos (o métodos, en la programación orientada a objetos) que ofrece cierta biblioteca para ser utilizado por otro software como una capa de abstracción.

Son usadas generalmente en las bibliotecas de programación.

Twitter pone a disposición de los usuarios tres APIs distintas. Dos son las que llaman "REST API"[1] y la otra es una "Streaming API". Las aplicaciones basadas en Twitter podrán usar las tres APIs distintas, combinadas para llevar a cabo sus objetivos.

El uso de las APIs de Twitter está limitado, por lo que tus aplicaciones no pueden conectarse un número indeterminado de veces para solicitar una operación. Sin embargo, los límites serían más o menos aceptables para páginas personales y proyectos pequeños (además, siempre podemos cachear los resultados para no tener que solicitar lo mismo muchas veces seguidas). En el caso que se desee construir sistemas que hagan un uso intensivo de la API de Twitter, estaría la posibilidad de registrar la aplicación. Los límites de acceso a la API sin registro son 150 solicitudes por hora, mientras que para aplicaciones registradas en la "whitelist" podrían llegarse a hacer 20.000 solicitudes por hora.

Ahora bien, las solicitudes a veces se cuentan dependiendo de la IP del sistema que conecta con Twitter y a veces dependiendo de la cuenta de usuario Twitter que está solicitando un recurso, por lo que estos límites podrían ser un poco mayores si tenemos uno o varios usuarios Twitter. Hay también otros límites de uso de la API, basados en los límites de paginación de las búsquedas que se pueden realizar, es decir, que limitan los resultados de búsquedas de estatus de un usuario o las referencias de una palabra clave en los post públicos.

Para comunicar con la API de Twitter, hay que trabajar en UTF-8 y enviar cualquier parámetro codificado con formato de URL. Estos formatos y juegos de caracteres se pueden conseguir fácilmente con diversas funciones de PHP [8].

### 4.1.1. API de Twitter y cURL

La API de Twitter funciona por HTTP, accediendo a URLs que devuelven datos, todo por el protocolo HTTP. Para facilitar la solicitud de URLs por parte de un cliente y procesar las

[1] REST API es una API web que funciona por HTTP y se accede a partir de URLs que devuelven contenidos en formatos distintos, como XML, JSON, HTML, etc.

respuestas del servidor a esas solicitudes, muy habitualmente se utiliza una librería aparte llamada cURL, que tiene una serie de funciones y procedimientos para acceder al contenido de URLs.

Para acceder a cURL se puede utilizar la línea de comandos, si es que el comando cURL está instalado en nuestro sistema, pero también se puede utilizar las funciones para cURL que tiene PHP.

Tal como indican en la propia documentación de la API de cURL, se podría lanzar este comando para acceder al "public timeline" (los últimos mensajes públicos enviados por todos los usuarios de Twitter):

*curl http://api.Twitter.com/1/statuses/public_timeline.rss*

Dado que este comando sirve para acceder a información pública, no requiere autenticación de un usuario para usar la API de Twitter. Hay muchas otras URL de consulta a la API de Twitter que sí requieren el login de un usuario [8].

## 4.2. Procedimiento de captura de datos

Primeramente, con el fin de conseguir los datos necesarios para la consecución de la clasificación automática, se ha decidido crear una cuenta de Twitter nueva y real (@cgonzalezTFG) para poder seguir así a los candidatos de los principales partidos políticos, y analizar sus tweets de una manera más personal.

Durante varias semanas, se ha realizado la recopilación de tweets, así como su almacenamiento, usando como punto de partida la campaña electoral andaluza.

La captura de datos se divide en dos fases:

- Fase de entrenamiento. Se realiza una recopilación de tweets (300) de opinión de las distintas cuentas, citadas más adelante, con el objetivo de fijar las reglas que posteriormente se usarán para la implementación del clasificador.
- Fase de validación. En paralelo con la fase de entrenamiento, se realiza una recopilación masiva de tweets para la validación del clasificador ya implementado.

Los principales partidos políticos y sus máximos representantes que se siguen en la cuenta @cgonzalezTFG, son los que se muestran a continuación en las diferentes tablas.

Ya que el Trabajo de Fin de Grado se comenzó en la campaña electoral andaluza, las cuentas analizadas son, en su mayor parte, de la comunidad de Andalucía. También se realiza un seguimiento de las cuentas de los principales partidos políticos a nivel nacional, así como de sus principales representantes.

| A NIVEL DE PARTIDO | | |
|---|---|---|
| **PSOE** | **PP** | **IU** |
| **@psoedeandalucia** | @ppandaluz | @iuandalucia |
| **#psoeandalucia** | @vota_ppandaluz | #iuandalucia |
| **#psoedeandalucia** | #PrimeroAndalucia | @iujaen |
| **#socialistasandaluces** | #PrioridadAndalucía | @iucordoba |
| **@psoejaen** | #ppandaluz | @iusevilla |
| **#psoejaén** | #StopSusana | @iuhuelva |
| **@psoecordoba** | #JuanmaSí | @iucadiz |
| **#psoecórdoba** | @pp_jaen | @iumalaga |
| **@PSOEdeSevilla** | #97Ilusiones | @iugranada |
| **#psoesevilla** | @pp_cordoba | @iualmeria |
| **@psoedehuelva** | @PPdeSevilla | #LaFuerzaDeLaIzquierda |
| **#psoeHuelva** | @Populareshuelva | |
| **@psoedecadiz** | @ppcadiz | |
| **#psoecadiz** | @PPMalaga | |
| **@psoemalaga** | @ppgranada | |
| **#psoemalaga** | @PP_Almeria | |
| **@psoegranada** | @ppmadrid | |
| **#psoegranada** | | |
| **@psoealmeria** | | |
| **#psoealmeria** | | |
| **@psoe** | | |

Tabla 1: Partidos políticos (1 de 2).

| A NIVEL DE PARTIDO | | |
|---|---|---|
| PODEMOS | **CIUDADANOS** | **UPyD** |
| **@Podemos_AND** | @Cs_Andalucia | @UPyD_Andalucia |
| **#ElCambioEmpiezaEnAndalucía** | @Cs_Jaen | @upyd_jaen |
| **@PodemosCordoba** | @Cs_Cordoba | @upydcordoba |
| **@PodemosJaen** | @Cs_Sevilla_ | @upydsevilla |
| **@PodemosSevilla** | @cs_huelva | @upyd_huelva |
| **@Podemos_Huelva** | @cs_cadiz | @upydcadiz |
| **@PodemosCadiz** | @CsMalaga | @upydmalaga |
| **@PodemosMalaga** | @csgranada | @upydgranada |
| **@PodemosGranada** | @Almeria_Cs | @upyd_almeria |
| **@PodemosAlmeria** | #AndalucíaPideCambio | #LevantaAndalucía |
| **@ahorapodemos** | @ciudadanoscs | |

Tabla 2: Partidos políticos (2 de 2).

| **A nivel de candidato** | | |
|---|---|---|
| PSOE | PP | **IU** |
| Micaela Navarro Garzón @micaela_navarro | Miguel Ángel García Anguita @GarciaAnguita | Juan Serrano Jódar @juanserranoiu |
| Juan Pablo Durán Sánchez @jpduran | María del Rosario Alarcón Mañas @rampppp | Elena Cortés Jiménez @ElenaCortesIU |
| Susana Díaz Pacheco @_susanadiaz | Juan Francisco Bueno Navarro @juanbuenopp | Antonio Maillo Cañadas @MailloAntonio |
| Mario Jesús Jiménez Díaz @mariojimenez | Manuel Andrés González Rivera @Magonzalezlepe | Rafael Sanchez Rufo @rafasrufo |
| María Teresa Jiménez Vílchez @teresajimenez64 | Ana María Maestre @AnaMestrePP | Inmaculada Nieto Castro @InmaNietoC |
| | Juan Manuel Moreno Bonilla @JuanMa_Moreno | Jose Antonio Castro Román @jacastro1974 |
| | Carlos Rojas Garcías @CarlosRojas_PPA | María del Carmen Pérez Rodríguez @MCarmenPerezIU |
| | María Carmen Crespo Díaz @CarmenCrespoPP | Rosalía Marín Escobar @rosalia_martin |
| Pedro Sanchez @sanchezcastejon | Mariano Rajoy @marianorajoy | |

**Tabla 3: A nivel de candidato (1 de 2).**

| **A nivel de candidato** | | |
|---|---|---|
| PODEMOS | **CIUDADANOS** | **UPyD** |
| Mercedes Barranco Rodríguez @Mer_barranco | María Isabel Albás Vives @isabelalbas | Miguel Ángel Garrido Jurado @garridojurado1 |
| David Jesús Moscoso Sánchez @_davidmoscoso | Juan Antonio Marín Lozano @juanmarin_cs | Rosario de la Haba @rosariohaba |
| Begoña María Gutiérrez Valero @BegoPodemos | Julio Jesús Díaz Robledo @juliojdiaztw | Martín Jacobo De La Herrán Sabick @mdlherran |
| María Teresa Rodriguez-Rubio Vázquez @teresarodr_ | Sergio Romero Jiménez @sergioromeroj | Patricia Pelegrín @ppelegrin |
| Félix Gil Sánchez @FelixGilPodemos | Irene Rivera Andrés @_Irene_Rivera | Carlos Márquez @UpydCarlos |
| Jose Luis Serrano Moreno @serranojoseluis | José Antonio Funes Arjona @joseafunesCs | Sergio Ocaña @SOcanaUPyD |
| Lucia Ayala Asensio @Lucia_Andalucia | Marta Bosquet Aznar @martabosquet | Ploma Meidna Rivas @pmedinarivas |
| | | Desiderio Enciso @Desiderio78 |
| Pablo Iglesias @Pablo_Iglesias_ | Albert Rivera @Albert_Rivera | |

**Tabla 4: A nivel de candidato (2 de 2).**

Para la obtención del conjunto de tweets que vamos a clasificar finalmente simplemente se llama a la API de Twitter desde PHP, usando la librería tmhoauth.

Para la fase de entrenamiento se obtienen un total de 300 tweets etiquetados a mano. Del mismo modo y mientras se realiza la fase de entrenamiento, de forma simultanea se recogen 1200 tweets más para la realización de la fase de validación.

# 5. Fase de entrenamiento

Durante esta primera fase se lleva a cabo la definición de las reglas y la posterior implementación del clasificador.

El motor de análisis que se va a utilizar para la clasificación, es el motor de reglas (actualmente en beta) del servicio MeaningCloud.


Figura 20: Empresa MeaningCloud [7].

## 5.1. Definición de reglas

Con el fin de elaborar una serie de reglas que posteriormente serán la base de la implementación del clasificador, se realiza una toma de datos de un grupo de tweets reducido, 300.

El motor de reglas ejecuta modelos de clasificación definidos en dos ficheros.

En el primer fichero se definen los siguientes parámetros:

- El **título del modelo**.
- La **descripción del modelo**, con una breve explicación.
- Los **parámetros de entrada**, que actúan como una macro que se puede usar en las reglas y cuyo valor cambia en cada llamada. Se entiende un parámetro por línea.
- Se puede aplicar filtros en función del:
    - **Número de categorías**: donde se pueden mostrar un número definido de categorías, por ejemplo si el modelo se clasifican en 5 categorías es posible mostrar las tres primeras, con la sentencia CAT_NUMBER = 3.Por defecto se muestran todos los resultados.
    - **Número de categorías con distinta relevancia**: se muestra como en el anterior caso, un número definido de categorías pero con la distinción que si la categoría cuarta tiene la misma importancia que la tercera, se mostrarán hasta la cuarta categoría con la sentencia REL_NUMBER = 3. Como en el caso anterior, por defecto se mostrarán todos los resultados.
    - **Relevancia absoluta**: con la sentencia REL_ABSOLUTE = n, se filtrarán las categorías con una relevancia absoluta inferior a la indicada (n). Por defecto n=0.

- **Relevancia relativa**: con la sentencia REL_RELATIVE = n, siendo n un número decimal del cero al uno, se filtrarán las categorías con una relevancia relativa a la primera, inferior al porcentaje indicado por n.

El segundo de los ficheros contiene las reglas de clasificación se pueden definir:

- **Macros**: se define una macro con una expresión, que se quiera sustituir por otra expresión bajo la responsabilidad del usuario que sea correcta. La sentencia utilizada en este caso sería: {MACRO} = EXPRESIÓN. Por ejemplo: {PYMES} = Pequeñas y medianas empresas.
- **Reglas de expresión**: se puede activar o desactivar clases, es decir, si se cumple la EXPRESIÓN, se suma un peso (WEIGHT) a la relevancia de esa clase.
  - **Si WEIGHT =** - , entonces la clase se elimina. Es como si tendiera a menos infinito (-∞).
  - **Si WEIGHT** = + , entonces la clase se incluye y además es la que mayor relevancia tiene. Como si tendiera a más infinito (+∞).
  - **Si WEIGHT =** ±**número** (positivo o negativo), se suma o se resta el peso de la clase.
  - **Si WEIGHT** no existe, por defecto se considera como un signo +.

## 5.2. Sintaxis de las reglas

Según los términos, se puede englobar las reglas en 5 grupos principalmente:

- Términos muy simples (VST: Very Simple Term), que pueden ser a su vez:
  - Término (Term) o palabra
  - Forma: F@term
  - Lema: L@term
  - A la forma/lema se le puede añadir una etiqueta: term@tag, donde la etiqueta puede ser un nombre (N), un verbo (V), un adjetivo (A) o un adverbio (E).
  - O añadir una etiqueta a la forma y al lema a la vez: L@term@tag
  - Añadir información semántica: S@sementity_type@sementity_class
  - Si ya se ha definido en otra clase: #CLASS
  - Si aparece la etiqueta solo: @tag

```
%%%%%%%%%%%%%%%%%%%%%%%%
%        Apoyo          %
%%%%%%%%%%%%%%%%%%%%%%%%
"L@todo con" -> #Apoyo
"con S@Top>Person" -> #Apoyo
"conS@Top>Person" -> #Apoyo
"con S@Top>Id" -> #Apoyo
apoyando|apoy[oa]|apoyos|apoyamos -> #Apoyo
apoyar -> #Apoyo
respaldo -> #Apoyo
```

**Figura 21: Ejemplo de etiquetado para la clase *Apoyo*.**

En la Figura 21, se define la clase *Apoyo*. Si en el tweet a clasificar aparece por ejemplo, "Con @PSOE", se clasificará como *Apoyo,* debido a que una regla coincide con la forma de la etiqueta ("con S@Top>Id").

Si aparece la palabra "apoyo/a" se clasifica como *Apoyo* gracias a la sentencia "apoy [oa]" (regla que puede verse en la Figura 21). Un ejemplo de la clasificación para la clase *Apoyo*, puede verse en la Figura 22.

| Text | Classes |
|---|---|
| @PSOE @sanchezcastejon @patxilopez Gracias Sr. Sanchez por no apoyar la única ley buena del PP. Esto le pasara a Ud. factura en las urnas. | Agradecimiento (100%) Gracias<br>Apoyo (100%) apoyar |
| @ximopuig: los avances democráticos, en igualdad, y en libertades fueron con @PSOE y serán con @sanchezcastejon http://t.co/vyKxrTQOaN | Igualdad (100%) igualdad<br>Apoyo (100%) con\|@PSOE\|@sanchezcastejon |

**Figura 22: Ejemplo de clasificación para la clase *Apoyo*.**

- Términos simples (ST: Simple Term)
    - Podemos concatenar términos simples con el operador lógico OR "|", para poder clasificar un término u otro.

Un ejemplo de concatenación con el operador lógico OR, tomando como referencia la Figura 21, puede ser la sentencia "apoyando|apoy[oa]|apoyos|apoyamos -> #Apoyo"; que concatena varios términos: apoyando, apoyo/a, apoyos, apoyamos. Si uno de estos términos aparece en el tweet que estamos clasificando, se clasificará el tweet como *Apoyo*.

- Términos:
    - Los términos se pueden afirmar, negar o poner varias palabras seguidas entre comillas para indicar que clasifique según esas palabras en concreto.

Unas reglas que demuestran ejemplos para la afirmación o negación de términos son (Figura 23):

```
%%%%%%%%%%%%%%%%%%%%%%%%
%       Justicia       %
%%%%%%%%%%%%%%%%%%%%%%%%
justicia -> #Justicia
+justo -> #Justicia
-justo -> #Injusticia
```

**Figura 23: Ejemplo de afirmación y negación de términos.**

Con esta regla se puede definir una misma palabra, anteponiéndole el signo positivo o negativo, a dos clases totalmente distintas u opuestas. Es decir, justo negado (-justo) significa injusticia o lo que es lo mismo: no es justo (Injusticia). En la Figura 24 se puede observar un ejemplo de clasificación de tweets que se clasifican como *Justicia*.

| Text | Classes |
|---|---|
| @ppmadrid @EsperanzAguirre Cuántos hospitales dices que has privatizado?Esos kms d metro van a la villa olimpica o a la ciudd d la justicia? | Economía (100%) has privatizado<br>Justicia (100%) justicia |
| RT @CiudadanosCs: .@Albert_Rivera "Hay que despolitizar la Justicia. Queremos un Poder Judicial independiente." #RiveraConFederico | Justicia (100%) Justicia |

**Figura 24: Ejemplo de clasificación de *Justicia*.**

Las reglas que definen la clase *Cambio* se muestran en la Figura 25.

```
%%%%%%%%%%%%%%%%%%%%%%
%       Cambio       %
%%%%%%%%%%%%%%%%%%%%%%
cambio|cambiado|cambiando -> #Cambio
cambiar -> #Cambio
"España necesita (un)? cambio" -> #Cambio
"Es necesario un cambio" -> #Cambio
```

**Figura 25: Reglas para la clase *Cambio*.**

Un ejemplo para la clasificación de términos consecutivos (o multiwords) y exactas se muestra a continuación (Figura 26):

| Text | Classes |
|---|---|
| @MonederoJC @Pablo_Iglesias_ @pnique Nota @JyPSpain Es necesario un cambio en la política migratoria... http://t.co/HPkmQPE3h2 | Cambio (100%) cambio |
| @Pablo_Iglesias_ Lo q queremos son personas comprometidas q cambien estas políticas y hagan leyes para q los corruptos paguen y desaparezcan | Corrupción (100%) corruptos<br>Cambio (100%) cambien |

**Figura 26: Ejemplo de clasificación de *Cambio*.**

Debido a este ítem cuando aparezca “Es necesario un cambio” se clasificará en la clase *Cambio.*

- O simplemente se pondrán términos para apuntar a una clase concreta (Véase ejemplo Figura 28).

```
%%%%%%%%%%%%%%%%%%%%%%%%
%   Agradecimiento     %
%%%%%%%%%%%%%%%%%%%%%%%%
gracias -> #Agradecimiento
gratitud -> #Agradecimiento
Gracias -> #Agradecimiento
agradecer -> #Agradecimiento
```

**Figura 27: Reglas para la clase *Agradecimiento*.**

| Text | Classes |
|---|---|
| RT @anarosaq: @marianorajoy @sanchezcastejon @Albert_Rivera @Pablo_Iglesias_ Muchas gracias por dejarme compartir vuestro tiempo #24horas. | Agradecimiento (100%) gracias |
| @ahorapodemos un grafico muy bueno, muchas gracias | Agradecimiento (100%) gracias |
| @ahorapodemos que piensa la gente de vuestro partido sobre videos como este? gracias https://t.co/uXk7JbPWCM | Agradecimiento (100%) gracias |

**Figura 28: Ejemplo de clasificación de *Agradecimiento*.**

Una combinación de varias reglas anteriores puede resultar de ayuda a la hora de clasificar. En la Figura 25 podemos ver la siguiente regla:

"España necesita (un)? cambio" -> #Cambio

Las comillas indican que queremos encontrar la concatenación de términos exacta, para que se clasifique en la clase *Cambio* (->#Cambio). Sin embargo, el término entre paréntesis acabado en interrogación puede existir o no.

| Text | Classes |
|---|---|
| RT @PSOEAlcaladeH: Ya queda menos para volver a tener un gobierno socialista, porque España necesita un cambio. Arriba @sanchezcastejon htt… | Necesidad (100%) necesita<br>Cambio (100%) España\|necesita\|un\|cambio |

**Figura 29: Ejemplo de clasificación para la clase *Cambio*.**

Como podemos observar en la Figura 29, el tweet contiene el término "un" y lo clasifica como *Cambio*, sin embargo, si no existe el término "un" también lo clasificaría como *Cambio* con la misma regla.

- Términos complejos
  - Término AND término: n términos enlazados con el operador lógico AND. Implica que tienen que aparecer los n términos en el tweet, sin importar el orden. Un ejemplo de tweet para esta regla se ha definido para la clase *Empleo* (Figura 30):

```
%%%%%%%%%%%%%%%%%%%%%%%%
%        Empleo          %
%%%%%%%%%%%%%%%%%%%%%%%%
empleo -> #Empleo
trabajo -> #Empleo
paro -> #Empleo
trabajo AND inem AND puestos -> #Empleo
```

**Figura 30: Reglas para la clase *Empleo*.**

| Text | Classes |
|---|---|
| colas inmensas en el inem, ¿realmente se están generando puestos de trabajo? | Empleo (100%) inem\|puestos\|trabajo |

**Figura 31: Ejemplo de clasificación anidando con el operador lógico AND.**

Es similar a la concatenación de términos con el operador lógico OR, con la principal diferencia de que si no existe uno de los dos términos, en este caso, no se clasificaría como *Empleo* (Figura 31).

  - También se pueden usar paréntesis para indicar la preferencia de los operadores.

- Si los tweets contienen caracteres especiales que necesitan ser clasificados, se ponen delante de estos caracteres la barra de división invertida: \.

| | | | |
|---|---|---|---|
| \+ | \( | \) | * |

Figura 32: Caracteres especiales que necesitan escape.

En el ámbito de la política, no tiene sentido clasificar estos caracteres, sin embargo, en el ámbito de la programación, cuando se habla de lenguajes de programación por ejemplo, sí existe la necesidad de "escapar" estos caracteres especiales para que sean clasificados correctamente. En la Figura 33 se propone un ejemplo de tweet que se clasificaría en este caso como *Programación*.

C\+\+ -> #Programación

| Text | Classes |
|---|---|
| Analista programador C++,Cen Madrid. Analista programador con experiencia en programación en C++, C, uso de bases de datos SQL y sistema operativo Linux | Programación (100%) C++ |

Figura 33: Ejemplo de clasificación con caracteres especiales.

## 5.3. Implementación del clasificador

Se copian los tweets a un Excel, para su mejor manejo, y se procede a la implementación del clasificador. Se ha definido para ello las etiquetas, es decir las palabras claves (definidas como clases) que se repiten en los tweets. Se consideran dos grandes grupos de etiquetas: Conceptos de política en sí, y conceptos que afectan directamente a los ciudadanos.

A la hora de implementar el clasificador, a pesar de tratarse de una tarea muy costosa, se optó por clasificar y etiquetar manualmente todos los tweets empleados para el entrenamiento del clasificador. Cabe destacar que el número total de tweets para esta fase de entrenamiento es mucho menor (el 20% del total de tweets recopilados) que los tweets que se utilizarán para la fase de validación.

A modo de resumen, en la Tabla 5 y en la Tabla 6, se pueden ver los conceptos que se consideran más importantes (etiquetas o clases), debido a su repetición, en las cuentas que seguimos en Twitter.

**Apoyo = "Un acto de apoyo a alguien"**
**Cambio = "Cuando se habla de un cambio en la política"**
**Éxito = "Se habla del éxito que está teniendo un acontecimiento electoral"**
**Agradecimiento = "Cuando se agradece a alguien algún gesto"**
**Debate = "Cuando hay un debate televisivo o una reunión en la que consideramos que se va a tratar temas políticos"**
**Elecciones = "Cuando se refieren a las elecciones"**

| |
|---|
| **Proyecto = "Desarrollo de un proyecto político"** |
| **Candidato = "Mención de un candidato a la presidencia"** |

**Tabla 5: Aspectos que atañen a los partidos políticos.**

| |
|---|
| **Empleo = "Cuando habla de empleo"** |
| **Justicia = "Cuando se habla de justicia"** |
| **Empresas = "Se habla de distintas empresas, pymes o autónomos"** |
| **Desigualdad = "Desigualdad de oportunidades"** |
| **Igualdad = "Se habla de igualdad de oportunidades"** |
| **Exclusión = "Cuando se habla de exclusión social"** |
| **Corrupción = "Cuando se habla de un partido o un candidato corrupto"** |
| **Esperanza = "Cuando se tiene esperanza en conseguir algún proyecto"** |
| **Inversión = "Cuando se realiza una inversión económica en algún tipo de proyecto"** |
| **Beca = "Cuando se habla de una beca de estudios"** |
| **Solidaridad = "Cuando hay un acto solidario"** |
| **Necesidad = "Cuando se habla de suplir una necesidad"** |
| **Desilusión = "Cuando se muestra el desencanto por algún motivo"** |
| **Economía = "Cuando se habla de la economía"** |

**Tabla 6: Aspectos que atañen a los ciudadanos.**

En el apartado Sintaxis de las reglas, se han visto algunas de las definiciones de reglas para la clasificación. A continuación se mostraran las definiciones de reglas que faltan por aparecer.

```
%%%%%%%%%%%%%%%%%%%%%%%%
%       Necesidad      %
%%%%%%%%%%%%%%%%%%%%%%%%
necesidad -> #Necesidad
necesitamos-> #Necesidad
necesita-> #Necesidad
```

```
%%%%%%%%%%%%%%%%%%%%%%%%
%       Empresas       %
%%%%%%%%%%%%%%%%%%%%%%%%
pymes|autónom[osas] -> #Empresas
```

```
%%%%%%%%%%%%%%%%%%%%%%%%
%       Proyecto       %
%%%%%%%%%%%%%%%%%%%%%%%%
proyecto -> #Proyecto
proyectos -> #Proyecto
```

```
%%%%%%%%%%%%%%%%%%%%%%%%
%       Candidato      %
%%%%%%%%%%%%%%%%%%%%%%%%
candidat[oa] -> #Candidato
candidat[osas] -> #Candidato
```

```
%%%%%%%%%%%%%%%%%%%%%%%%
%     Corrupción       %
%%%%%%%%%%%%%%%%%%%%%%%%
corrupción|corrupt[osas] -> #Corrupción
imputad[osas] -> #Corrupción
imputad[oa] -> #Corrupción
```

```
%%%%%%%%%%%%%%%%%%%%%%%%
%     Igualdad        %
%%%%%%%%%%%%%%%%%%%%%%%%
igualdad|igualdades -> #Igualdad
```

```
%%%%%%%%%%%%%%%%%%%%%%%%
%      Exclusión       %
%%%%%%%%%%%%%%%%%%%%%%%%
exclusión|excluid[oa]|excluidos|excluidas -> #Exclusión
```

```
%%%%%%%%%%%%%%%%%%%%%%%%
%    Desigualdad       %
%%%%%%%%%%%%%%%%%%%%%%%%
desigualdad|desigualdades -> #Desigualdad
violencia|género -> #Desigualdad
violencia AND "de género" -> #Desigualdad
```

```
%%%%%%%%%%%%%%%%%%%%%%%%
%        Debate         %
%%%%%%%%%%%%%%%%%%%%%%%%
debate|debates-> #Debate
debatir-> #Debate
```

```
%%%%%%%%%%%%%%%%%%%%%%%%
%      Esperanza        %
%%%%%%%%%%%%%%%%%%%%%%%%
esperanza|esperar|esperamos -> #Esperanza
ilusión|ilusiones -> #Esperanza
ilusionado[s]|ilusionada[s] -> #Esperanza
```

```
%%%%%%%%%%%%%%%%%%%%%%%%
%        Inversión      %
%%%%%%%%%%%%%%%%%%%%%%%%
inversión -> #Inversión
inversiones -> #Inversión
```

```
%%%%%%%%%%%%%%%%%%%%%%%%
%     Solidaridad       %
%%%%%%%%%%%%%%%%%%%%%%%%
solidario|solidaria -> #Solidaridad
solidarios|solidarias -> #Solidaridad
solidaridad -> #Solidaridad
```

```
%%%%%%%%%%%%%%%%%%%%%%%%
%        Economía       %
%%%%%%%%%%%%%%%%%%%%%%%%
economía -> #Economía
económico -> #Economía
económica -> #Economía
privatizar -> #Economía
pensiones -> #Economía
```

```
%%%%%%%%%%%%%%%%%%%%%%%%
%     Elecciones        %
%%%%%%%%%%%%%%%%%%%%%%%%
Elecciones -> #Elecciones
"campaña electoral" -> #Elecciones
```

```
%%%%%%%%%%%%%%%%%%%%%%%%
%         Beca          %
%%%%%%%%%%%%%%%%%%%%%%%%
beca|becas-> #Beca
```

```
%%%%%%%%%%%%%%%%%%%%%%%%
%      Desilusión       %
%%%%%%%%%%%%%%%%%%%%%%%%
desencanto -> #Desilusión
desilusionado -> #Desilusión
desilusionada -> #Desilusión
```

```
%%%%%%%%%%%%%%%%%%%%%%%%
%         Éxito         %
%%%%%%%%%%%%%%%%%%%%%%%%
éxito|exitosa -> #Éxito
ganar -> #Éxito
ganando -> #Éxito
ganador -> #Éxito
ganadora -> #Éxito
```

**Tabla 7: Algunas definiciones de reglas.**

# 6. Resultados

Como ya se ha mencionado en varias ocasiones a lo largo del Trabajo Fin de Grado, el número total de tweets recopilados es de 1500. De éstos, 300 han sido utilizados en la fase de entrenamiento y los 1200 restantes son los que se han utilizado en esta fase para la validación del clasificador.

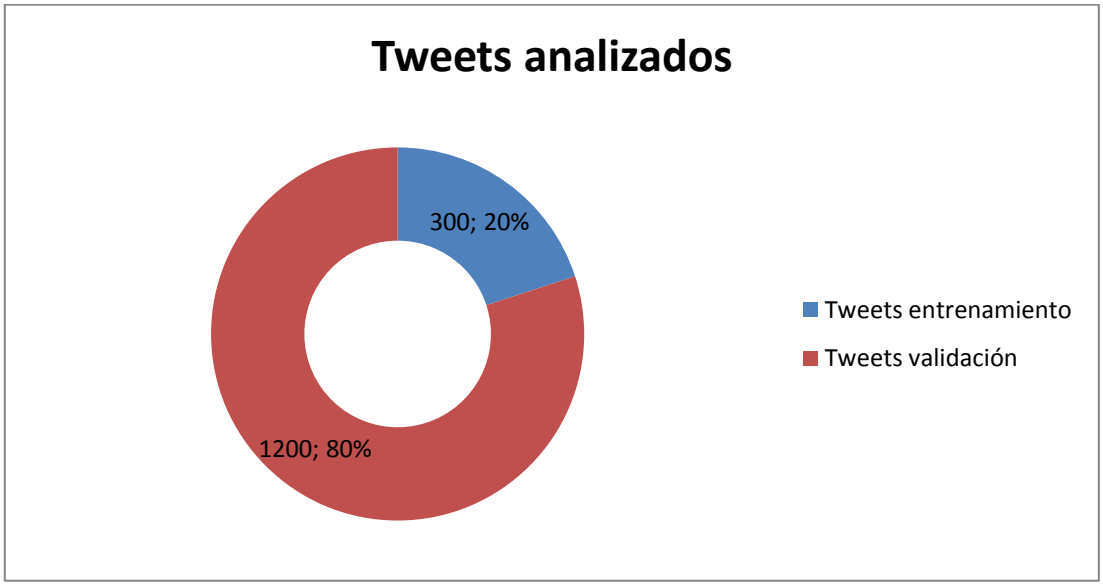



Figura 34: Tweets analizados.

La evaluación del clasificador constituye la etapa final del Trabajo Fin de Grado. Dado un conjunto de tweets para clasificar, se obtendrá una respuesta del clasificador, ordenando los tweets en cada una de las categorías. Para saber si la clasificación ha sido satisfactoria, es necesario realizar una evaluación de los resultados.

La matriz de confusión, o también llamada tabla de contingencia, es una matriz cuadrada nxn, donde n es el número de clases identificadas. El número de instancias clasificadas correctamente es la suma de la diagonal de la matriz y el resto están clasificadas de forma incorrecta. Las filas de la matriz representan los valores de predicción para el modelo, mientras que las columnas representan los valores reales. Comprobando el número de elementos no nulos fuera de la diagonal principal de la matriz, se podrá obtener una buena aproximación de la calidad del clasificador.

Los parámetros de exactitud para cada clase son los siguientes:

- **True Positive (TP)** es el total de tweets que fueron correctamente clasificados como clase x, es decir, qué cantidad de la clase ha sido capturada.
  En la matriz de confusión, es el valor del elemento de la diagonal, es decir, el valor en el que confluyen cada fila con la columna de la misma clase.

- **False Positive (FP)** es la proporción de tweets que fueron clasificados como clase x, pero en realidad pertenecen a otra clase, de entre todos los tweets que no tienen clase x.
  En la matriz de confusión, es la suma de la columna menos el valor del elemento de la diagonal (TP).
- **False Negative (FN)** es la proporción de tweets que no han sido clasificados como tal.
  En la matriz de confusión, es la suma de la fila menos el valor del elemento de la diagonal (TP).
- **True Negative (TN)** es la proporción de tweets que no pertenecen a la clase y no han sido clasificados en ella.
- **Precisión** es la proporción de tweets que de veras tienen clase x entre todos los que fueron clasificados como clase x.
  En la matriz es el elemento de la diagonal dividido por la suma de la columna relevante.

$$precision = \frac{tp}{tp + fp}$$

Ecuación 1: Precisión.

- **Recall (o cobertura)** representa la cobertura del clasificador, es decir, la cantidad de tweets que clasifica frente a los no clasificados y clasificados. Un sistema puede clasificar todos los tweets en una categoría, aunque lo haga mal, teniendo pues una cobertura alta pero una precisión baja.

$$recall = \frac{tp}{tp + fn}$$

Ecuación 2: Recall (o cobertura).

- **F-Measure (o medida-F)** La situación ideal es aquella en la que existe una precisión y cobertura alta (es decir muy cercana a 1). A esta situación se la denomina utilidad teórica. Con el objeto de ponderar y ver cuán lejos están ambas medidas de la utilidad teórica, suelen emplearse los valores de ambas métricas combinadas en una media armónica denominada medida-F.

$$medida - F = \frac{(1 + \beta^2) * precision * recall}{(\beta^2 * precision) + recall}$$

Ecuación 3: Medida-F.

Donde β es un parámetro que permite estimar la influencia relativa de ambas medidas: precisión y cobertura. Si se considera proporcionar igual peso a ambas características (β = 1) la medida final a considerar que determinará las prestaciones del clasificador será la siguiente:

$$medida - F = \frac{2 * precision * recall}{precision + recall}$$

Ecuación 4: Medida-F considerando β =1.

Otros parámetros que se podrían tomar en cuenta, a la hora de validar el clasificador son:

- **Velocidad**: La velocidad de ejecución hoy día es uno de los factores más importantes. En sistemas donde se tienen que clasificar cientos de noticias al segundo, de nada sirve la precisión o la cobertura si el sistema es lento. En este trabajo no lo tendremos en cuenta, ya que no es un objetivo definido en él.
- **Claridad**: Las reglas que permiten al sistema realizar la clasificación deben ser simples y sencillas. La claridad de las reglas son claras y precisas como se ha visto en algunos ejemplos en el etiquetado de clases.

Para obtener las estimaciones de la precisión y la cobertura se pueden emplear dos métodos:

- ***Micro-averaging***: la precisión y la cobertura se obtienen sumando todas las decisiones individuales.

- ***Macro-averaging***: la precisión y la cobertura se evalúan en primer lugar de forma local para cada categoría, y después se hace la media con los resultados para las diferentes categorías.

Es decir, mientras que el micro-averaging cada documento recibe igual peso (es una medida centrada en el documento), en el macro-averaging las categorías reciben igual peso (medida centrada en la categoría).

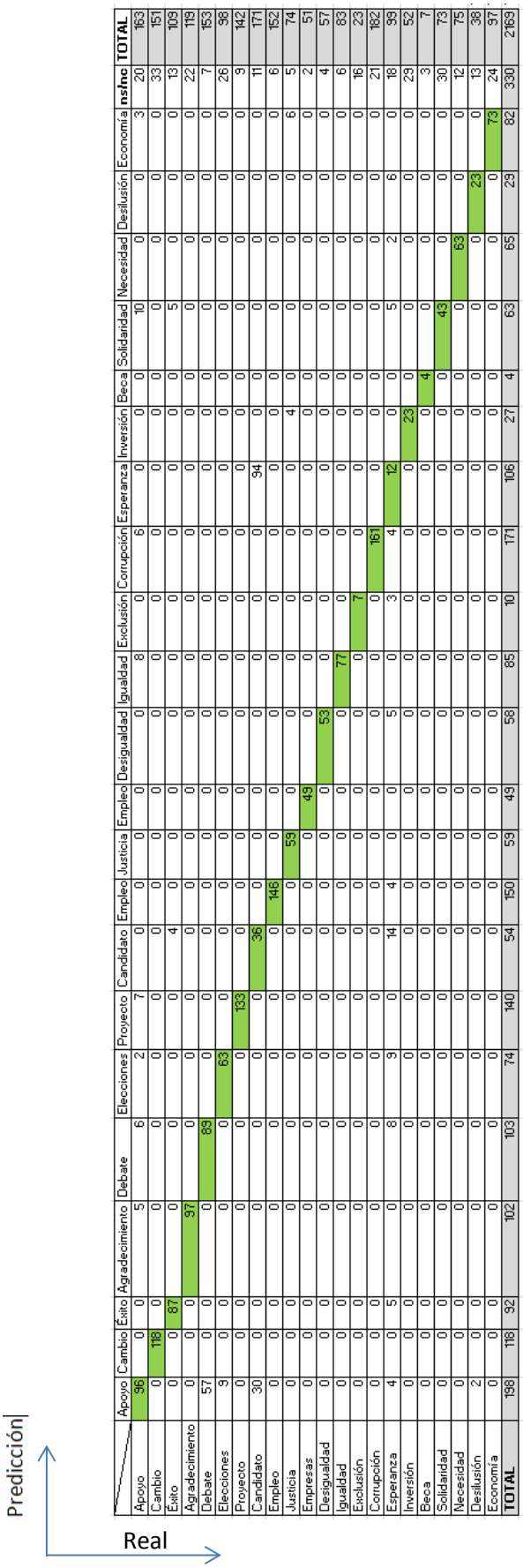
En la Figura 35 puede verse a modo de resumen, una matriz de confusión 2x2 donde están los primeros cuatro parámetros definidos anteriormente.

| | | Valor en la realidad | | |
|---|---|---|---|---|
| | | ***p*** | ***n*** | **total** |
| **Predicción outcome** | ***p'*** | Verdaderos Positivos | Falsos Positivos | P' |
| | ***n'*** | Falsos Negativos | Verdaderos Negativos | N' |
| | **total** | P | N | |

Figura 35: Matriz confusión 2x2.

## 6.1. Matriz de confusión

Tras los resultados obtenidos de la clasificación de los 1200 tweets, se construye la matriz de confusión (Tabla 8) con la cual posteriormente analizaremos los parámetros de exactitud vistos en el apartado anterior.




| | Apoyo | Cambio | Éxito | Agradecimiento | Debate | Elecciones | Proyecto | Candidato | Empleo | Justicia | Empleo | Desigualdad | Igualdad | Exclusión | Corrupción | Esperanza | Inversión | Beca | Solidaridad | Necesidad | Desilusión | Economía | nsnc | TOTAL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Apoyo | 96 | 0 | 0 | 0 | 5 | 6 | 2 | 7 | 0 | 0 | 0 | 0 | 8 | 0 | 6 | 0 | 0 | 0 | 10 | 0 | 0 | 3 | 20 | 163 |
| Cambio | 0 | 118 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 33 | 151 |
| Éxito | 0 | 0 | 87 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 13 | 109 |
| Agradecimiento | 0 | 0 | 0 | 97 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 22 | 119 |
| Debate | 57 | 0 | 0 | 0 | 89 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 153 |
| Elecciones | 9 | 0 | 0 | 0 | 0 | 63 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 26 | 98 |
| Proyecto | 0 | 0 | 0 | 0 | 0 | 0 | 133 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 | 142 |
| Candidato | 30 | 0 | 0 | 0 | 0 | 0 | 0 | 36 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 94 | 0 | 0 | 0 | 0 | 0 | 0 | 11 | 171 |
| Empleo | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 146 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 152 |
| Justicia | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 59 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 6 | 5 | 74 |
| Empresas | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 43 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 51 |
| Desigualdad | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 53 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 57 |
| Igualdad | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 77 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 83 |
| Exclusión | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 16 | 23 |
| Corrupción | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 161 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 21 | 182 |
| Esperanza | 4 | 0 | 5 | 0 | 8 | 3 | 0 | 14 | 4 | 0 | 0 | 5 | 0 | 3 | 4 | 12 | 0 | 0 | 5 | 2 | 6 | 0 | 18 | 99 |
| Inversión | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 23 | 0 | 0 | 0 | 0 | 0 | 29 | 52 |
| Beca | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 3 | 7 |
| Solidaridad | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 43 | 0 | 0 | 0 | 30 | 73 |
| Necesidad | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 63 | 0 | 0 | 12 | 75 |
| Desilusión | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 23 | 0 | 13 | 38 |
| Economía | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 73 | 24 | 97 |
| **TOTAL** | 198 | 118 | 92 | 102 | 103 | 74 | 140 | 54 | 150 | 59 | 43 | 58 | 85 | 10 | 171 | 106 | 27 | 4 | 63 | 65 | 29 | 82 | 330 | 2169 |

**Tabla 8. Matriz de confusión**

Como se puede observar en la matriz, se definieron 22 clases (descritas en la Tabla 5 y Tabla 6), que como se ha explicado anteriormente, son las palabras más representativas de los tweets de entrenamiento.

Cada casilla de la diagonal principal de la matriz de confusión (casillas en verde), coinciden con los valores de TP de la clase en cuestión. La suma total de las filas de la matriz indican los resultados reales de la clasificación, es decir, de 163 tweets (total de la primera fila de la matriz) de la clase *Apoyo,* 96 son clasificados correctamente, el resto el sistema predijo que 5 eran de la clase *Agradecimiento,* 6 de *Debate*, 2 de Elecciones, 7 de *Proyectos,* 8 de *Igualdad,* 6 de *Corrupción,* 10 de *Solidaridad* y 3 de *Economía*. Además, el sistema no ha dado ninguna categoría (NS/NC) a 20 tweets.

Siguiendo con la misma lógica 151 tweets son clasificados como *Cambio* (segunda fila de la matriz), siendo el total de tweets de esta clase 118 y los no categorizados 33.

A partir de la matriz de confusión se obtienen las métricas de evaluación por clase.

| Categorías | TP | FP | FN | PRECISIÓN | COBERTURA | MEDIDA-F |
|---|---|---|---|---|---|---|
| **Apoyo** | 96 | 102 | 67 | 48,48% | 58,90% | 53,19% |
| **Cambio** | 118 | 0 | 33 | 100,00% | 78,15% | 87,73% |
| **Éxito** | 87 | 5 | 22 | 94,57% | 79,82% | 86,57% |
| **Agradecimiento** | 97 | 5 | 22 | 95,10% | 81,51% | 87,78% |
| **Debate** | 89 | 14 | 64 | 86,41% | 58,17% | 69,53% |
| **#Elecciones** | 63 | 11 | 35 | 85,14% | 64,29% | 73,26% |
| **Proyecto** | 133 | 7 | 9 | 95,00% | 93,66% | 94,33% |
| **Candidato** | 36 | 18 | 135 | 66,67% | 21,05% | 32,00% |
| **Empleo** | 146 | 4 | 6 | 97,33% | 96,05% | 96,69% |
| **Justicia** | 59 | 0 | 15 | 100,00% | 79,73% | 88,72% |
| **Empresas** | 49 | 0 | 2 | 100,00% | 96,08% | 98,00% |
| **Desigualdad** | 53 | 5 | 4 | 91,38% | 92,98% | 92,17% |
| **Igualdad** | 77 | 8 | 6 | 90,59% | 92,77% | 91,67% |
| **Exclusión** | 7 | 3 | 16 | 70,00% | 30,43% | 42,42% |
| **Corrupción** | 161 | 10 | 21 | 94,15% | 88,46% | 91,22% |
| **Esperanza** | 12 | 94 | 87 | 11,32% | 12,12% | 11,71% |
| **Inversión** | 23 | 4 | 29 | 85,19% | 44,23% | 58,23% |
| **Beca** | 4 | 0 | 3 | 100,00% | 57,14% | 72,73% |
| **Solidaridad** | 43 | 20 | 30 | 68,25% | 58,90% | 63,24% |
| **Necesidad** | 63 | 2 | 12 | 96,92% | 84,00% | 90,00% |
| **Desilusión** | 23 | 6 | 15 | 79,31% | 60,53% | 68,66% |
| **Economía** | 73 | 9 | 24 | 89,02% | 75,26% | 81,56% |
| **Total** | 1512 | 327 | 657 | | | |
| **Macro-Averaging** | | | | 83,86% | 68,37% | 74,15% |
| **Micro-Averaging** | | | | 82,22% | 82,22% | 82,22% |

**Tabla 9: Tabla resumen de las características del clasificador.**

En la siguiente gráfica podemos ver con mayor facilidad los 1200 tweets clasificados, donde se comparan los TP de las 22 clases definidas.

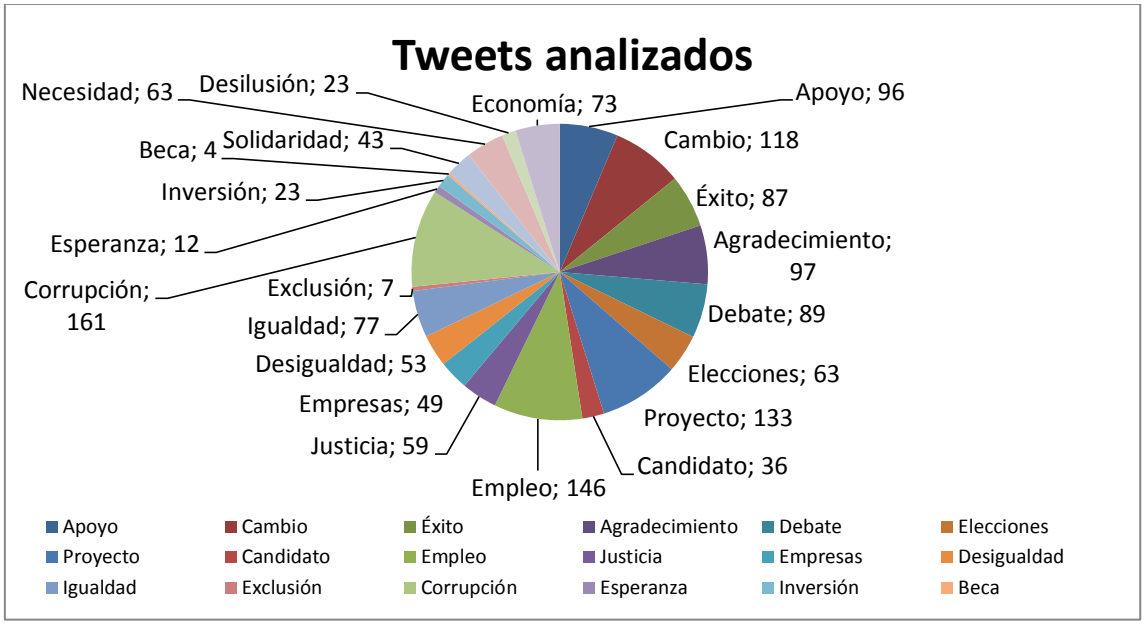



**Figura 36: TP tweets clasificados.**

Como podemos observar en la siguiente gráfica, más de la mitad de las clases tienen una cobertura mayor del 60%. Esto quiere decir que el clasificador es bastante aceptable ya que el número de tweets clasificados (de media) es mayor frente a los no clasificados.

Por el contrario nos encontramos con varias clases donde las coberturas tienen un menor porcentaje. Estas son "*Candidato*" y "*Esperanza*".

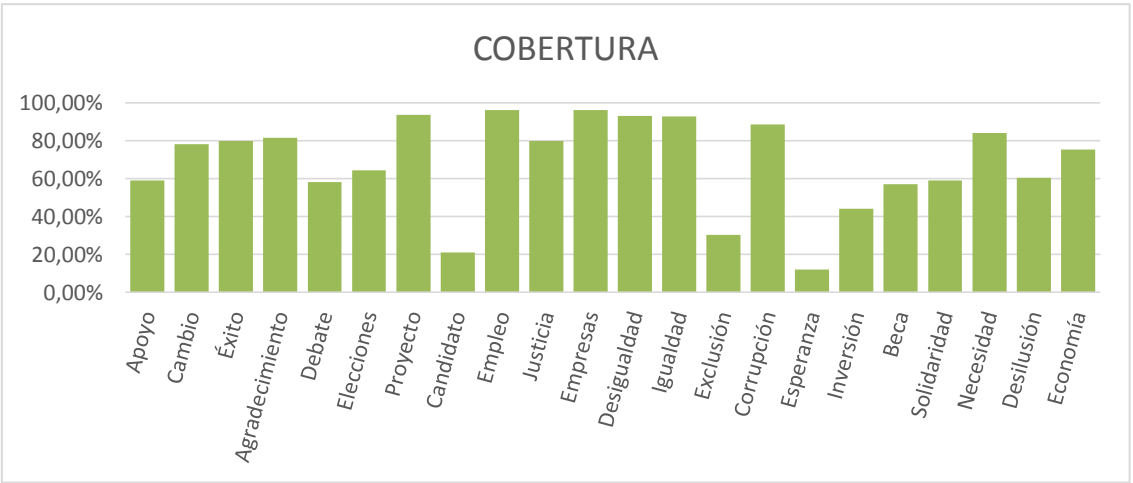



**Figura 37: Gráfica de cobertura del clasificador.**

Como podemos observar en la ecuación de precisión definida con anterioridad (Ecuación 1), el valor de la Precisión es menor cuanto mayor es el número de falsos positivos clasificados en la matriz.

Las dos clases que tienen menor porcentaje de precisión son *Esperanza* y *Apoyo.* A continuación se analiza cuáles son los motivos de estos valores bajos de Precisión.

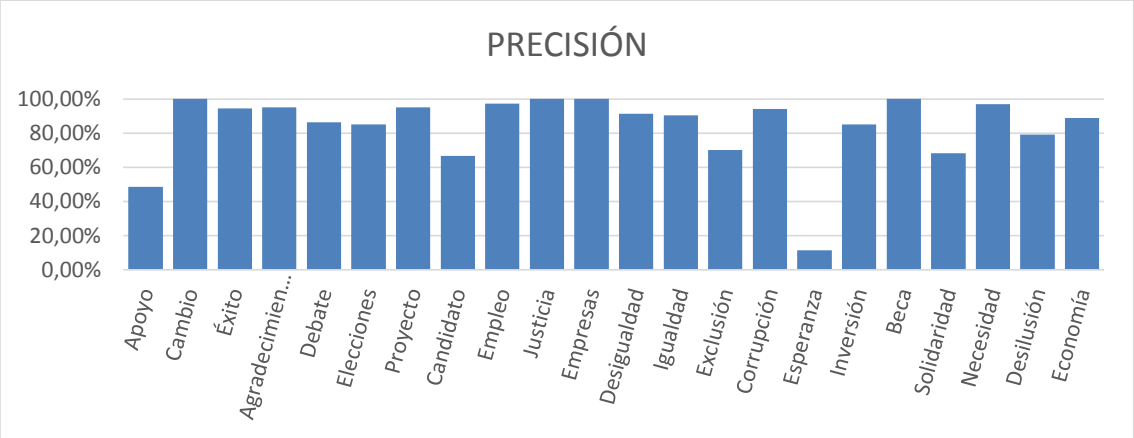



**Figura 38: Gráfica de precisión del clasificador.**

Muchos de los tweets que se clasifican como *Apoyo* son erróneos debido a que en el código, se consideraba como de la clase *Apoyo* cuando en el tweet te encontrabas con la expresión:

*"con S@Top>Person" -> #Apoyo.*

Muchos de los tweets y debido a que nos encontrábamos en un periodo de campaña electoral, hacían propaganda de cuando saldrían los candidatos en debates políticos en emisoras de radio o canales televisivos, con lo que contenían → con "nombre candidato", por lo que se clasifican mal, ya que no expresan apoyo a nadie.

Algo parecido ocurría con el caso de la clase *Esperanza*. Debido a que el nombre de esperanza es confundido en la mayoría de los tweets con el nombre de la presidenta del Partido Popular y portavoz del Grupo Municipal Popular en el Ayuntamiento de Madrid, Esperanza Aguirre. Dichos tweets como es evidente, no expresan un sentimiento de esperanza, sino que se dirigían expresamente a esta persona.

| TUIT | CLASS 1 |
|---|---|
| @ppmadrid @EsperanzAguirre jajaja, lo que puede salir de esa auditoría... me parto la caja con tus tweets, Esperanza | Esperanza |

**Figura 39: Ejemplo de error en la clasificación en clase Esperanza.**

Como se puede observar en la matriz de confusión (Tabla 8) muchos de los tweets que se clasificaron erróneamente como *Apoyo* en realidad se tendrían que haber clasificado como *Debate*. Muchos de estos tweets tenían el formato: "con (nombre candidato)", para anunciar el día y la hora que se produciría un debate televisivo o a través de la radio.

| TUIT | CLASS 1 |
|---|---|
| RT @PSpresidente: .@sanchezcastejon tendrá un encuentro este miércoles - 18h con @RicardoLagos, exPresidente de | Apoyo |

Figura 40: Ejemplo de error en la clasificación en clase Apoyo.

Otra gran mayoría de tweets que se deberían de haber clasificado como *Candidato*, se clasificaron erróneamente como *Apoyo*, debido a que muchos de ellos tenían el formato "con (nombre candidato)" comparando candidatos.

En cuanto a la clase *Solidaridad*, claramente tiene el mismo formato "con (nombre)" con lo que se clasifican erróneamente como *Apoyo*. Se muestra un ejemplo de la clase de *Solidaridad* que se clasificaron como *Apoyo*, Figura 41.

| TUIT | CLASS 1 |
|---|---|
| RT @ahorapodemos: Nos solidarizamos con @marianorajoy. Debe ser muy cansado memorizar todos los recortes que le manda Angela Merkel https:/… | Apoyo |

Figura 41: Ejemplo de error en la clasificación en clase Apoyo.

En el tweet del ejemplo (Figura 41) al encontrar "con @marianorajoy", se clasifica como *Apoyo*, aunque el autor del tweet se esté refiriendo de manera metafóricamente, eso sí, a "solidarizarse" con Mariano Rajoy por los recortes efectuados a los ciudadanos.

Casos similares a estos han llevado a clasificar también como *Apoyo* varios tweets que deberían haberse clasificado como *Igualdad, Proyecto, etc..*

Se puede apreciar que cuando los valores de FN y FP son parecidos, los porcentajes calculados de la cobertura y precisión son similares entre sí, debido a las ecuaciones vistas anteriormente (Ecuación 1: Precisión. y Ecuación 2: Recall (o cobertura).).

Se podría decir entonces que el sistema, ya que para el caso contrario parece también corroborarse, posee una relación entre la capacidad para clasificar tweets y cantidad de tweets que clasifica bien, es decir, cuanto mayor es la muestra, mejores resultados se obtienen con el clasificador, como se pueden ver en los resultados de macro y micro averaging al tener unas medidas similares (Tabla 9).

Se puede observar que no existe ninguna clase que sobresalga sobre el resto de clases. Sin embargo si se han definido varias clases que tienen valores de cobertura y precisión muy por debajo de la media.

Las clases definidas con mayor repercusión en la red social en la que se basa este estudio son *Cambio, Proyecto, Empleo y Corrupción,* como se puede observar ya que son las clases con mayor número de TP.

En cuanto a la precisión del clasificador respecto a las 22 clases definidas del conjunto total de tweets, se puede observar en la gráfica de precisión (Figura 38), que en prácticamente la totalidad de las clases se obtiene una precisión mayor al 80%.

En cuando a la Medida-F del clasificador, medida que relaciona precisión y cobertura, observamos que la mayoría de las clases (17) superan el 60% (Figura 42).

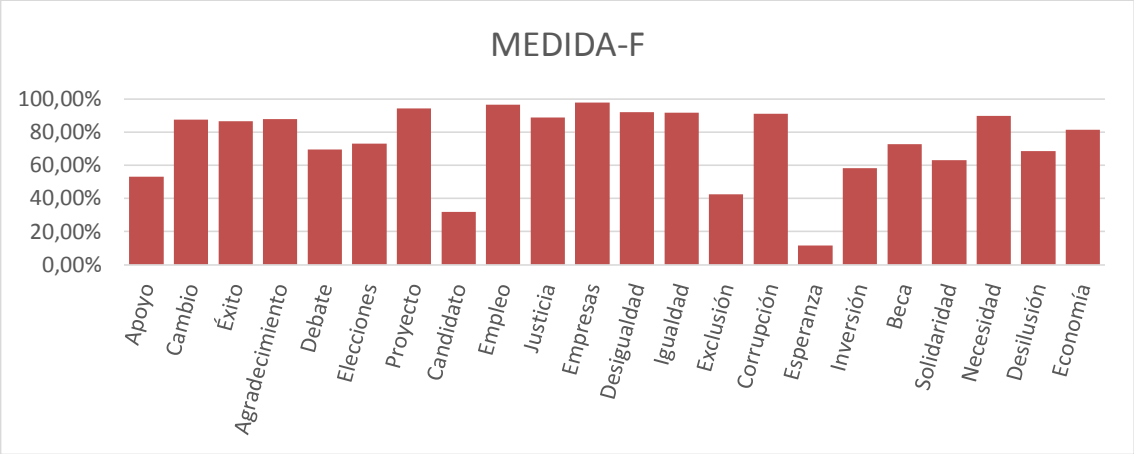



**Figura 42: Gráfica de Medida-F del clasificador.**

# 7. Presupuesto

Por último, dentro del apartado de presupuesto, vamos a realizar el cálculo real que ha tenido finalmente el trabajo, especificando las horas reales que se han empleado para realizarlo y los materiales que finalmente se han utilizado.

## 7.1. Descripción del proyecto

Autora: Cristina González Rubio

Departamento: Ingeniería Telemática.

Título: Clasificación automática de texto para el seguimiento de campañas electorales en redes sociales.

Duración: inicio 6 marzo de 2015 y finalización el 27 septiembre 2015. En total se ha invertido 600 horas en la realización del presente Trabajo Fin de Grado.

## 7.2. Planificación del Trabajo Fin de Grado

El Trabajo Fin de Grado consiste en la creación de unas reglas para la clasificación automática de texto y validación del clasificador con un grupo de tweets. Para el desarrollo del clasificador se han seguido las siguientes fases:

- Recopilación de información: En esta fase se ha llevado a cabo la recopilación de información relativa a clasificadores, así como de los algoritmos existentes de clasificación y métodos de aprendizaje. Como se clasifican tweets, también se ha recopilado información sobre las redes sociales que existen hasta el momento centrándonos en la red social Twitter. Por otro lado se ha recopilado información sobre la legalidad vigente respecto la privacidad del usuario.
- Implementación y recopilación de tweets: En esta fase se han obtenido los tweets con los cuales se han implementado las reglas del clasificador, y por otra parte se han obtenido tweets para posteriormente validar el clasificador.
- Validación del clasificador: Con los tweets obtenidos en la fase anterior, se procede a la validación del clasificador, donde se construye una tabla de confusión con las clases y sus características.
- Creación de la memoria.

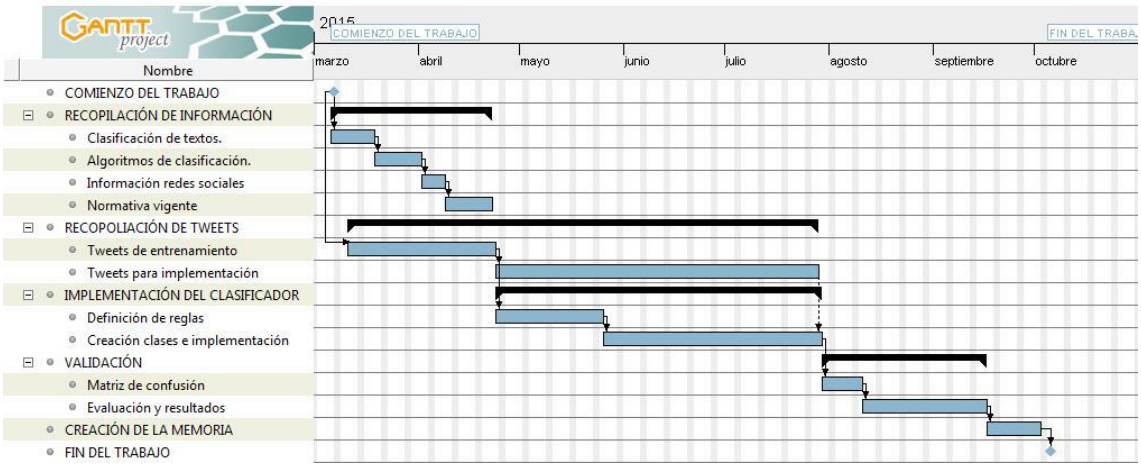



Figura 43: Diagrama de Gantt.

## 7.3. Cálculo de costes

### 7.3.1. Costes de personal

En el coste del personal se van a detallar las personas que han participado en el desarrollo del trabajo, la categoría, lo que cuesta a la hora cada persona, las horas que han trabajado y el coste total.

Se tendrá en cuenta los siguientes perfiles:

- Un ingeniero senior (Julio Villena Román): Tutor del Trabajo Fin de Grado. Realiza tareas de supervisión y consultor.
- Una ingeniero junior (Cristina González Rubio): Autora del Trabajo Fin de Grado.

| FASE | PERSONAL | COSTE/HORA | TOTAL HORAS | TOTAL COSTE |
|---|---|---|---|---|
| **Recopilación de información** | I.JUNIOR | 20 €/H | 136 H | 2720 € |
| | I.SENIOR | 42 €/H | 13 H | 546 € |
| **Diseño del clasificador** | I.JUNIOR | 20 €/H | 88 H | 1760 € |
| | I.SENIOR | 42 €/H | 8 H | 336 € |
| **Implementación del clasificador** | I.JUNIOR | 20 €/H | 188 H | 3760 € |
| | I.SENIOR | 42 €/H | 20 H | 840 € |
| **Validación del clasificador** | I.JUNIOR | 20 €/H | 140 H | 2800 € |
| | I.SENIOR | 42 €/H | 13 H | 546 € |
| **Creación de la memoria** | I.JUNIOR | 20 €/H | 48 H | 960 € |
| | I.SENIOR | 42 €/H | 4 H | 168 € |
| | | | TOTAL | 14436 € |

### 7.3.2. Coste de equipos

Para la realización de este proyecto se ha necesitado un equipo durante la duración total del mismo.

| Descripción | Coste | Dedicación | Periodo de amortización (3 años) |
|---|---|---|---|
| **Portatil Desarrollo (Toshiba Tecra S5 Core 2 Duo T7500 2.2 GHz. 4GB RAM. 250GB. Windows XP Professional)** | 1320 € | 28 semanas | 156 semanas |
| | | TOTAL | 236,92 € |

No se tiene en cuenta el coste de software, dado que no se ha necesitado ningún programa específico con coste adicional.

Programas utilizados:

- Paquete OpenOffice (tareas de ofimática)
- RefWorks (gestor de referencias)
- Ganttproject (planificación del trabajo)

### 7.3.3. Costes indirectos

Se estiman unos costes indirectos del 10% de los costes directos.

### 7.3.4. Costes totales

- Total costes directos: 14672,92 €
- Total costes indirectos: 1467,29 €

Total del Trabajo Fin de Grado: 16140,21 €

# 8. Conclusiones

En este estudio se realiza la obtención de datos de la red social Twitter a través de la API de Twitter. Se utiliza un primer grupo de tweets para implementar un clasificador automático de texto basado en el etiquetado de clases, centrado en el seguimiento de campañas electorales.

Tras la implementación, se utiliza un número mayoritario de tweets para su estudio y posterior validación.

Durante la fase de validación del clasificador se obtienen una serie de resultados significativos sobre el funcionamiento del clasificador.

Los resultados muestran que el clasificador presenta niveles de acierto buenos en la clasificación de mensajes de Twitter. Basándonos en lo expuesto en este apartado y en los datos vistos en el apartado de resultados, se puede concluir que el análisis del clasificador automático arroja unos resultados óptimos, ya que los valores de Medida-F (que combina la precisión y la cobertura del sistema) son prácticamente en su totalidad por encima del 60%.

Los sistemas actuales de clasificación basada en texto, aparte del algoritmo de aprendizaje, se apoyan además en distintos métodos basados en reglas, así como otros, para obtener mejores precisiones y eliminar posibles errores. El actual estudio ha demostrado que el factor semántico y la ambigüedad de las distintas categorías, son problemas más que importantes, ya que es difícil, incluso de manera manual, clasificar algunos tweets en una u otra categoría.

Cabe destacar que los tweets de entrenamiento, son distintos que los 1200 tweets que componen el cuerpo de clasificación. Debido a esto, algunas clases (*Apoyo, Esperanza, etc.*) se han desviado de su clasificación correcta, como se ha comentado anteriormente, ya que las hipótesis de clasificación planteadas durante la fase de entrenamiento no han resultado todo lo satisfactorias como cabía esperar en la clasificación definitiva.

Esta desviación también está provocada en parte, porque las opiniones de cada individuo pueden incluir ironía, sarcasmo, cinismo u otras figuras literarias que desvirtúan el sentido del tweet, haciendo extremadamente difícil su clasificación automática.

A la vista de los resultados se puede concluir que existe una necesidad de mejora a la hora de definir las reglas para las clases con menor cobertura, como por ejemplo Apoyo y Esperanza, ya que ha quedado demostrado que existen varios errores en su clasificación.

Tras la evaluación de todos los resultados se concluye que, aunque la definición de las reglas para algunas clases requieren de un estudio posterior para minimizar los datos más desfavorables, el funcionamiento general del clasificador automático basado en reglas que se valida en este Trabajo Fin de Grado es óptimo.

Por lo tanto se puede afirmar que se han cumplido con los objetivos planteados.

# 9. Trabajos futuros

Una vez finalizado el trabajo, se proponen a continuación varias posibles líneas de investigación sobre la clasificación automática, así como posibles vías de mejora para el clasificador desarrollado.

- Tal y como se ha revelado en las conclusiones de este trabajo, la primera idea de futuros trabajos es la mejora de las reglas de clases para las clases con peores resultados de clasificación.

- Ampliar el estudio al marco de las elecciones nacionales, revisando, modificando y/o ampliando las clases utilizadas, adaptándolas a las opiniones inherentes al nuevo marco de estudio. Los intereses políticos son muy diferentes dependiendo del marco en el que se realicen, por lo que es necesario revisar las reglas de clases si se cambia de periodo a analizar.

- Realizar un análisis de la evolución de los mensajes de uno o varios partidos políticos o de uno o varios representantes políticos durante un cambio en un periodo de tiempo determinado, por ejemplo, durante el cambio de campaña electoral a resultados de elecciones.

- Realizar un estudio similar al descrito en este trabajo centrando la clasificación de textos en dos de los partidos políticos mayoritarios. Analizar los resultados detallando las diferencias entre partidos políticos de diferentes ideologías y, por lo tanto, diferentes fines políticos.

- Analizar un periodo similar al estudiado en este trabajo en otra red social o blog político. Comparar los resultados obtenidos y los expuestos en este trabajo, analizando las diferencias existentes entre los clasificadores cuando las opiniones clasificadas están restringidas por un número reducido de caracteres y cuando no existe esta restricción.

# Anexo

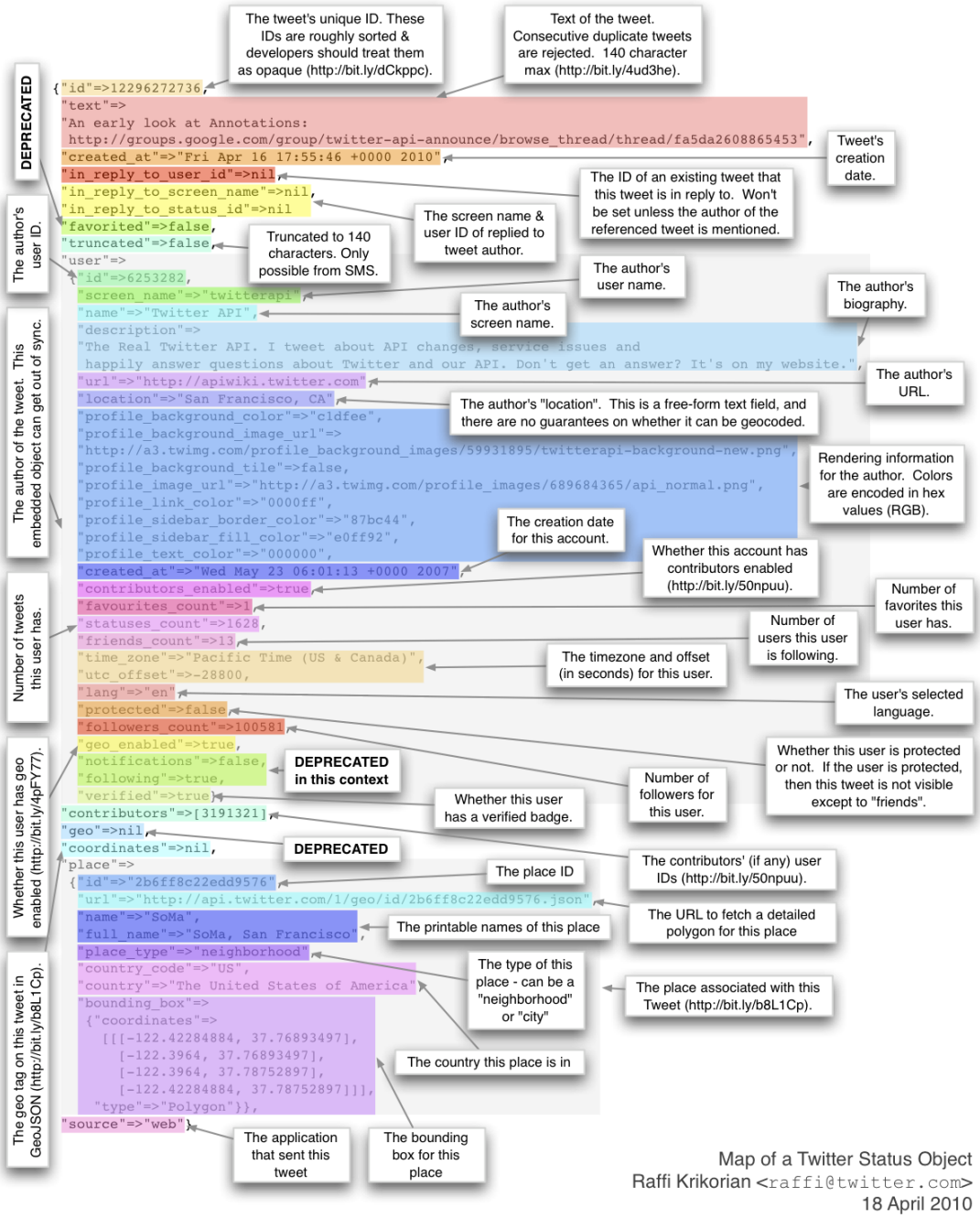



**Ilustración 1: Mapa de estado de un objeto de Twitter.**

# Bibliografía

[1]. 12 Herramientas Para Twitter: Gestión, Monitorización y Análisis. http://www.bluecaribu.com/8-herramientas-para-Twitter-gestion-monitorizacion-y-analisis

[2]. Agencia Española De Protección De Datos. http://www.agpd.es

[3]. Algoritmo, Definición. http://lema.rae.es/drae/?val=algoritmo;

[4]. Aprendizaje Supervisado y no Supervisado. http://redesneuronares.blogspot.com.es/

[5]. Árboles De Clasificación. http://iie.fing.edu.uy/ense/asign/recpat/material/tema3_00-01/node26.html

[6]. La Historia De Facebook, Paso a Paso. http://www.infotechnology.com/internet/La-historia-de-Facebook-paso-a-paso-20140203-0003.html

[7]. MeaningCloud API Service Level Agreement (“SLA”) http://www.Meaningcloud.Com.

[8]. Introducción a La API De Twitter. http://www.desarrolloweb.com/articulos/intro-api-Twitter-curl.html;

[9]. Ley Orgánica 15/1999, De 13 De Diciembre, De Protección De Datos De Carácter Personal.

[10]. Marketing De Contenidos: Todo Lo Que Debes Saber. http://www.bluecaribu.com/marketing-contenidos/

[11]. Redes Neuronales Artificiales. http://www.lab.inf.uc3m.es/~a0080630/redes-de-neuronas/

[12]. Las Redes Sociales Con Más Activos En 2015. http://goandweb.com/

[13]. Twitter Usage / Company Facts. https://about.Twitter.com/company

[14]. Winbold Data Systems. http://www.winbold.com/applied-analytics/machine-learning/

[15]. ALI, K.; MANGANARIS, S.and SRIKANT, R. Partial Classification using Association Rules. , 1997.

[16]. BUDDEEWONG, S.; and KREESURADEJ, W. A New Association Rule-Based Text Classifier Algorithm. Washington, DC, USA, ed. , 2005.

[17]. Carlos G. Figuerola, José L. Alonso Berrocal, Angel F. Zazo Rodríguez, Emilio Rodríguez. Algunas Técnicas De Clasificación Automática De Documentos.

[18]. Carlos G. Figuerola, José L. Alonso Berrocal, Angel F. Zazo Rodríguez, Emilio Rodríguez. Diseño De Un Motor De Recuperación De La Información Para Uso Experimental y Educativo.

[19]. CARMONA SUAREZ, E. Tutorial Sobre Máquinas De Vectores Soporte.

[20]. COLUMBICH, Dario. Interfaz De Programación De Aplicaciones (API).

[21]. Filippo Chieco; PÉREZ PÉREZ, Carlosand RODRÍGUEZ LUQUE, Joana. Algoritmo Del Vecino Más Cercano Aplicado a La web Filmaffinity.Com.

[22]. FRANCISCO, V.; and GERVÁS GÓMEZ-NAVARRO, P. Análisis De Dependencias Para La Marcación De Cuentos Con Emociones. Procesamiento Del Lenguaje Natural, ISSN 1135-5948, Nº. 37, 2006, 2006, pp. 137-144.

[23]. GARCÍA GALLO, B. Zapata Dimite Como Edil De Cultura Por El “dolor Generado” Por Sus Tuits.

[24]. HAYES, P. J.; ANDERSEN, P. M.and NIRENBURG G.I.B. SCHAMANDT, L. M. TCS: A Shell for a Contentbased Text Categorization. in Proceedings of CAIA-90, 6th IEEE Conference on Artificial Intelligence Applications, 1990, pp. 320-326.

[25]. HOPKINS, Jim. Surprise! there's a Third YouTube Co-Founder. USA Today.

[26]. Interactive Advertising Bureau. VI Estudio Redes Sociales.

[27]. LIU, B.; HSU, W.and MA, Y. Integrating Classification and Association Rule Mining . In KDD'98 ed. , 1998.

[28]. MARON, M. Automatic Indexing: An Experimental Inquiry, 1961, pp. 404--417.

[29]. MITCHELL, T. M. Machine Learning. New York US: McGraw Hill, 1996.

[30]. MORENO SECO, Francisco. Clasificadores Eficaces Basados En Algoritmos Rápidos De Búsqueda Del Vecino Más Cercano.

[31]. ORIHUELA, J. L. Mundo Twitter. Barcelona: Alienta, 2011.

[32]. PISCITELLI, A. Prólogo: Twitter, La Revolución y Los Enfoques Ni-Ni. En Orihuela, J.L., Mundo Twitter (Pp. 15-20). . Barcelona: Alienta, 2011.

[33]. PONCE-K IDATZIA, Isabel. Monográfico: Redes Sociales. Definición De Redes Sociales.

[34]. SANCHO CAPARRINI, Fernando. Introducción Al Aprendizaje Automático. http://www.cs.us.es/~fsancho/?e=75